FROM FAB TO TOKEN

THE STATE OF THE AI INFRASTRUCTURE MARKET

QCON AI BOSTON · APRIL 2026

JORDAN NANOS · MEMBER OF TECHNICAL STAFF · SEMIANALYSIS

boston.qcon.ai/presentation/boston2026/fab-token-state-market

semianalysis

INTRODUCTION

Who is SemiAnalysis?

SemiAnalysis is the best-of-breed Semiconductor and AI Research boutique. Our coverage spans the entire supply chain from semiconductor fabrication essentials to cutting-edge AI models, software, and the infrastructure that runs them.

Product-first approach

Analysts have extensive operating and research experience in the semiconductor and AI industry. We talk directly to engineers, buyers, and operators to understand the technology itself.

Connecting the dots

No other independent research house spends most of its time in the field seeking out public data points, applying technical expertise, and deriving actionable insights.

Whole-stack coverage

From wafers, CoWoS and HBM to networking, CPO, datacenter capacity, hyperscaler and neocloud server TCO to tokenomics, we work as one global team across the entire AI supply chain.

280,000+
newsletter subscribers

11
integrated industry models

6,000+
data centers tracked globally

100+
conferences attended annually

WHAT IS THE CHOKEPOINT?

This talk covers four things the market cares about. Critically, we assume models and talent are not the bottleneck, they keep improving.

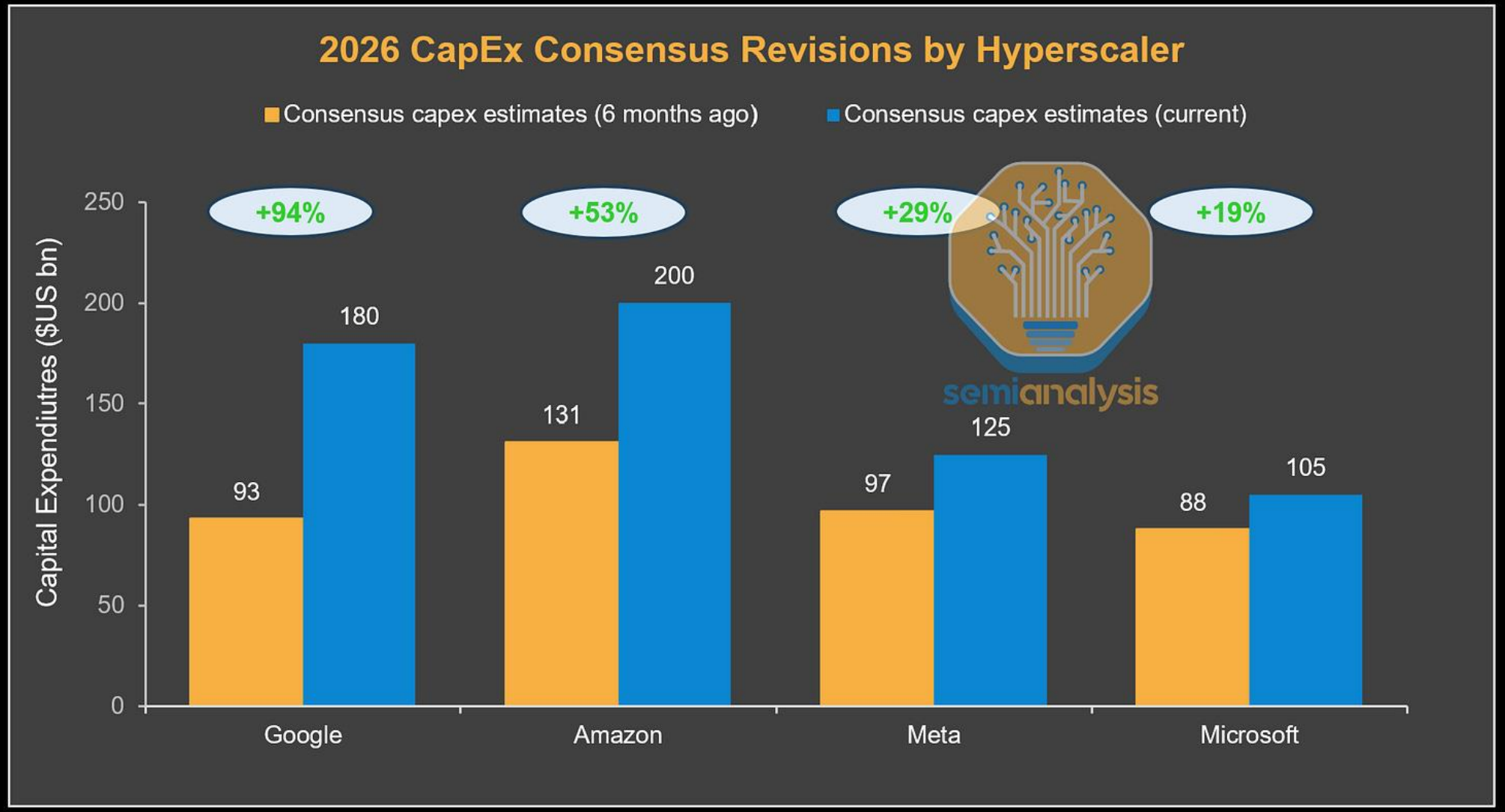
01	**Chips**	CoWoS, HBM, and wafer starts restrict how many accelerators reach customers each quarter.
02	**Datacenters**	Multi-GW campuses with mandatory liquid cooling, 800VDC, and behind-the-meter gas turbines.
03	**System Performance**	ClusterMAX for goodput and productivity, InferenceX for accelerators and their impact on margins.
04	**End User Demand**	Tokenomics: ChatGPT MAU growth, Claude Code vs Codex, open-source impact, free vs paid monetization, and API revenue.

PART 01

Chips

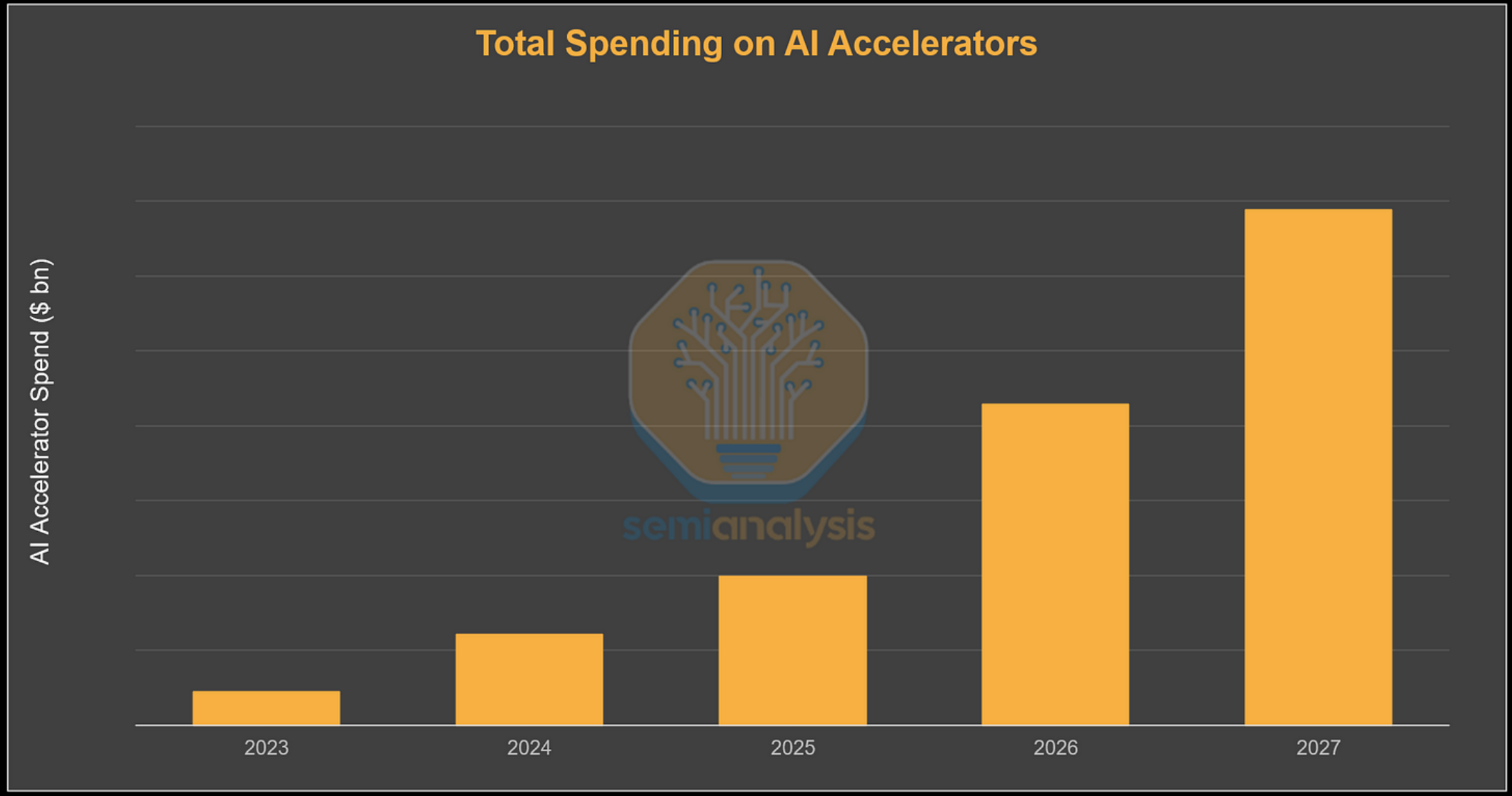
The accelerator market is gated by advanced packaging and HBM

HYPERSCALERS ARE SPENDING A LOT OF MONEY ON CHIPS

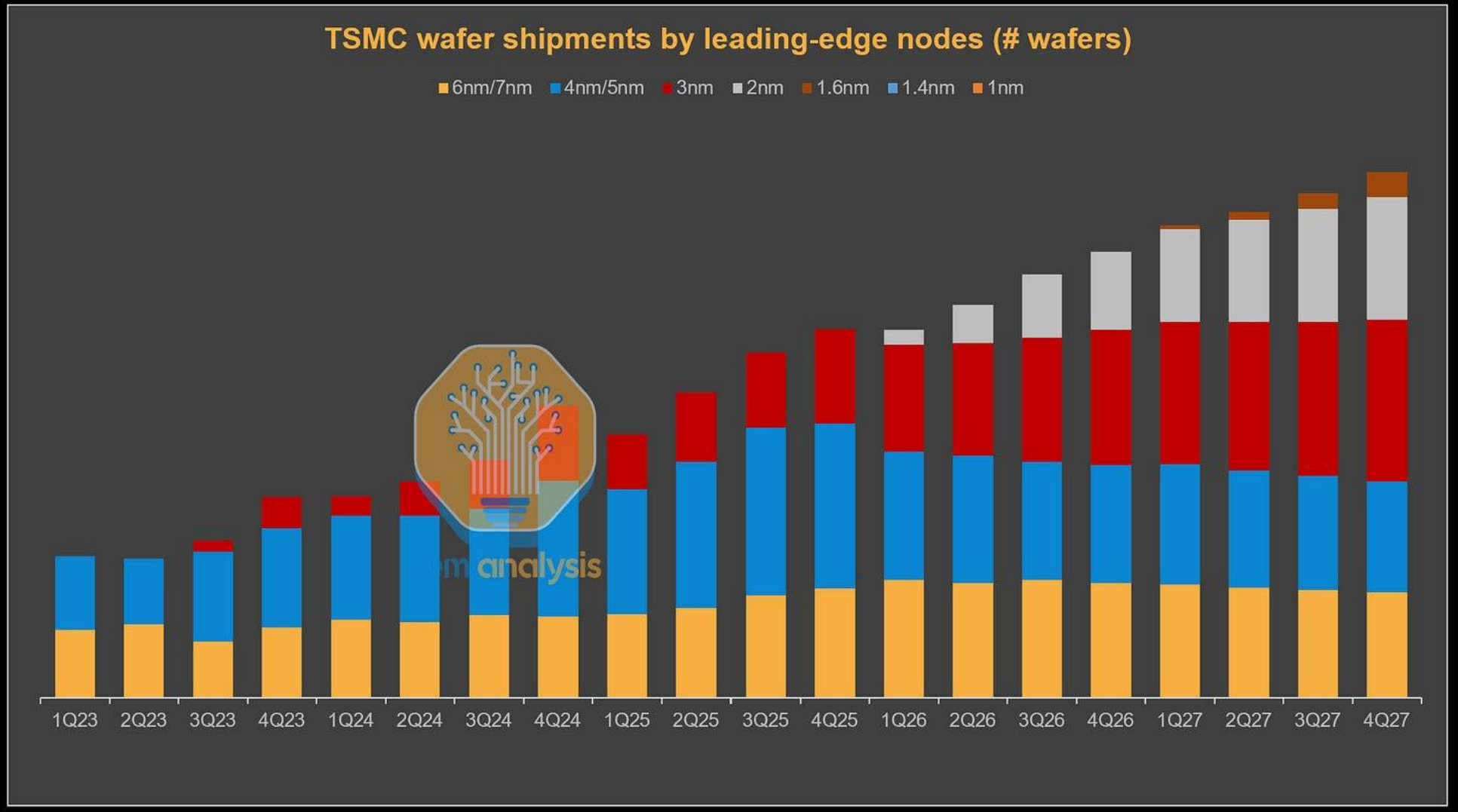


*Microsoft capex does not include capital leases.

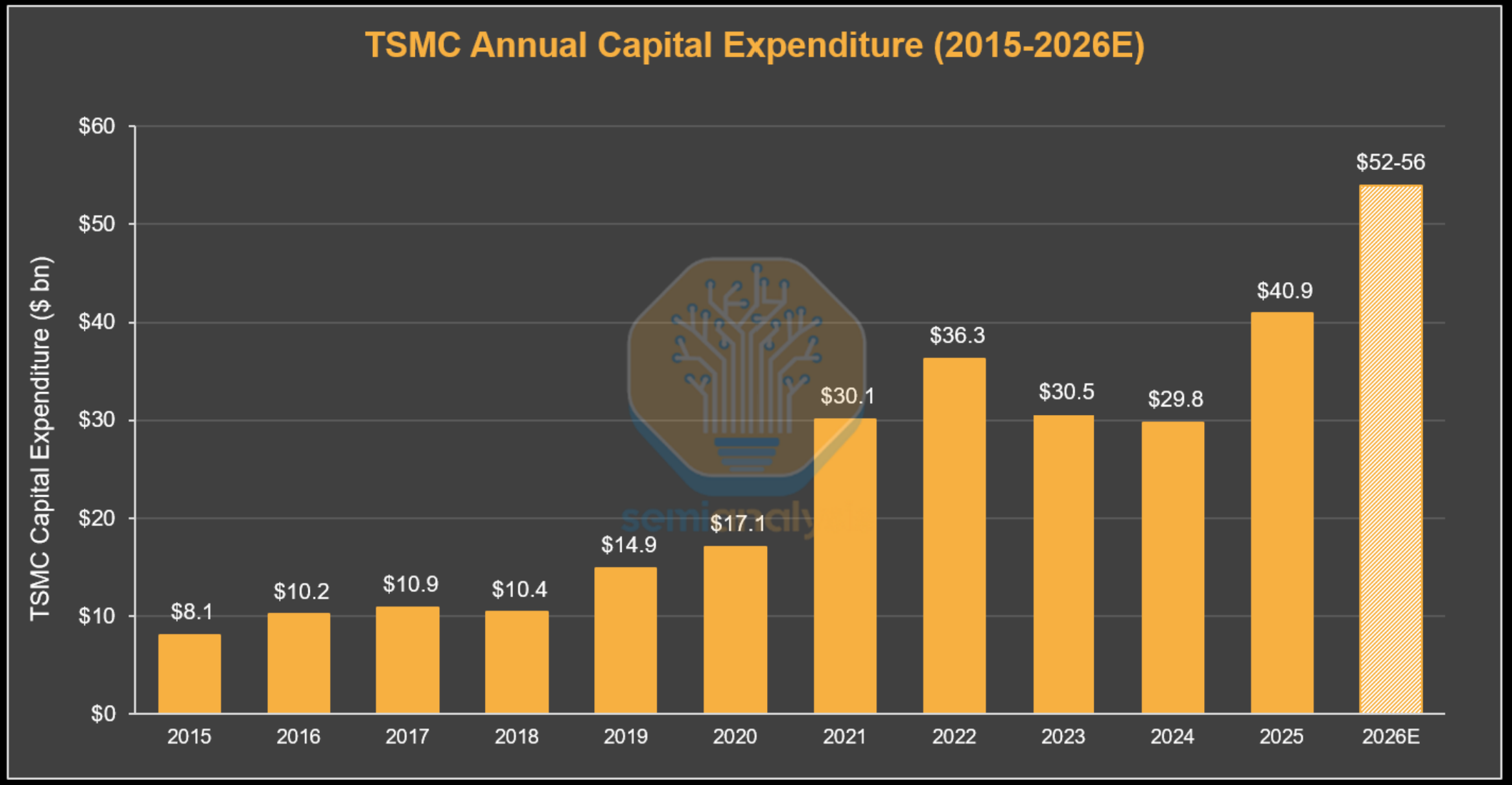
HYPERSCALERS ARE SPENDING A LOT OF MONEY ON CHIPS



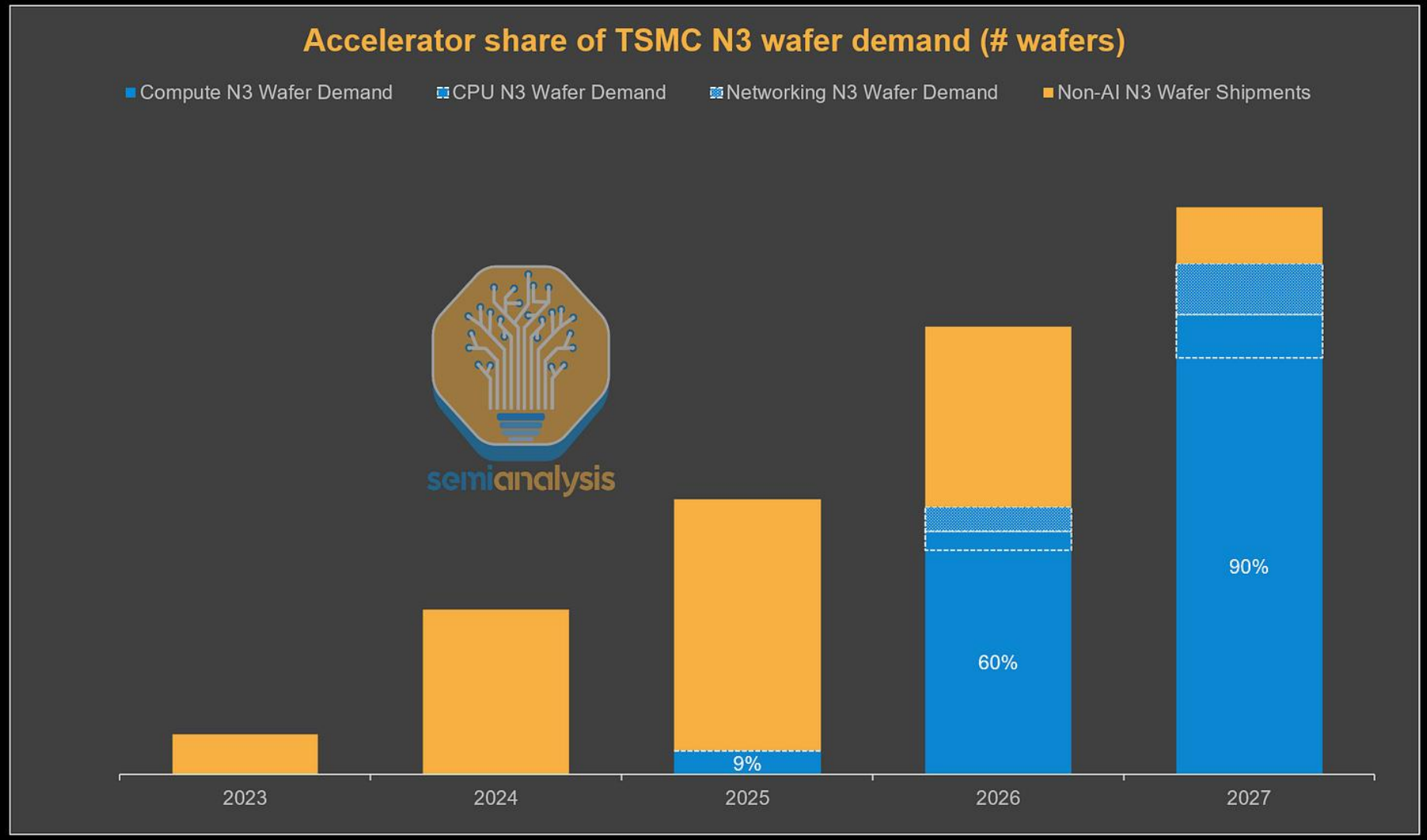
TSMC IS NOT KEEPING UP WITH DEMAND GROWTH



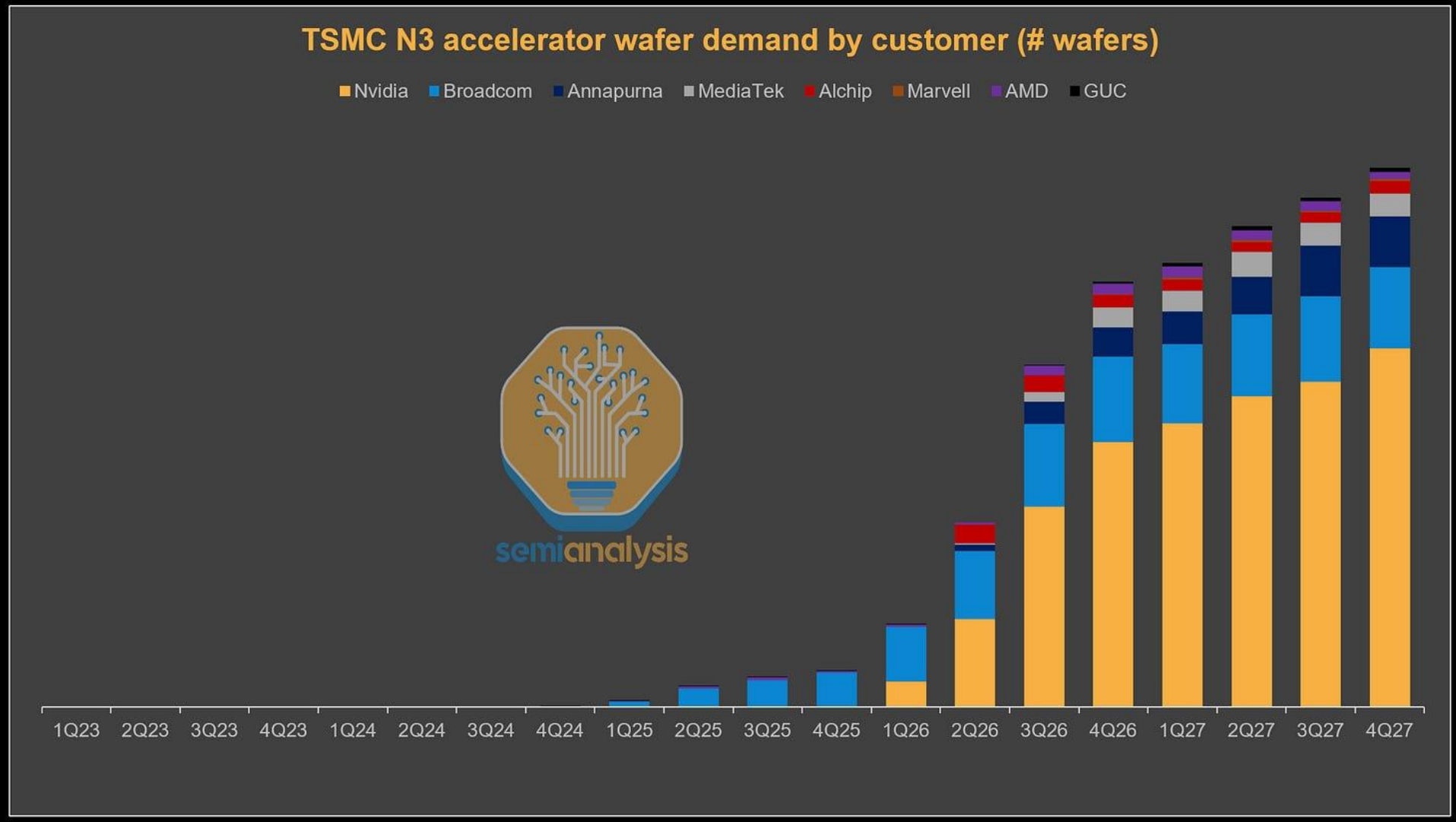
TSMC IS NOT KEEPING UP WITH DEMAND GROWTH



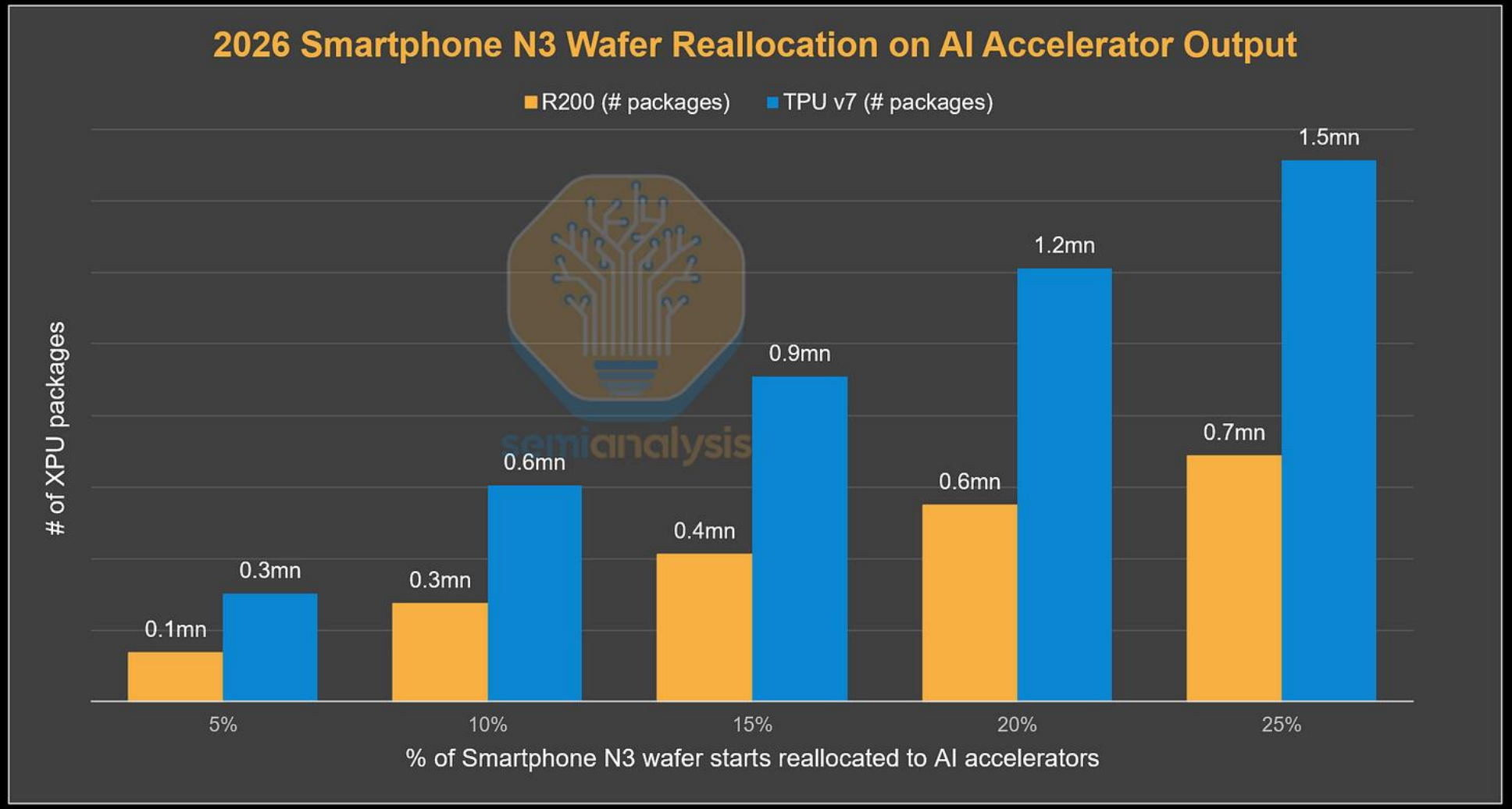
TO BE CLEAR, ALL OF THE GROWTH IS BECAUSE OF AI



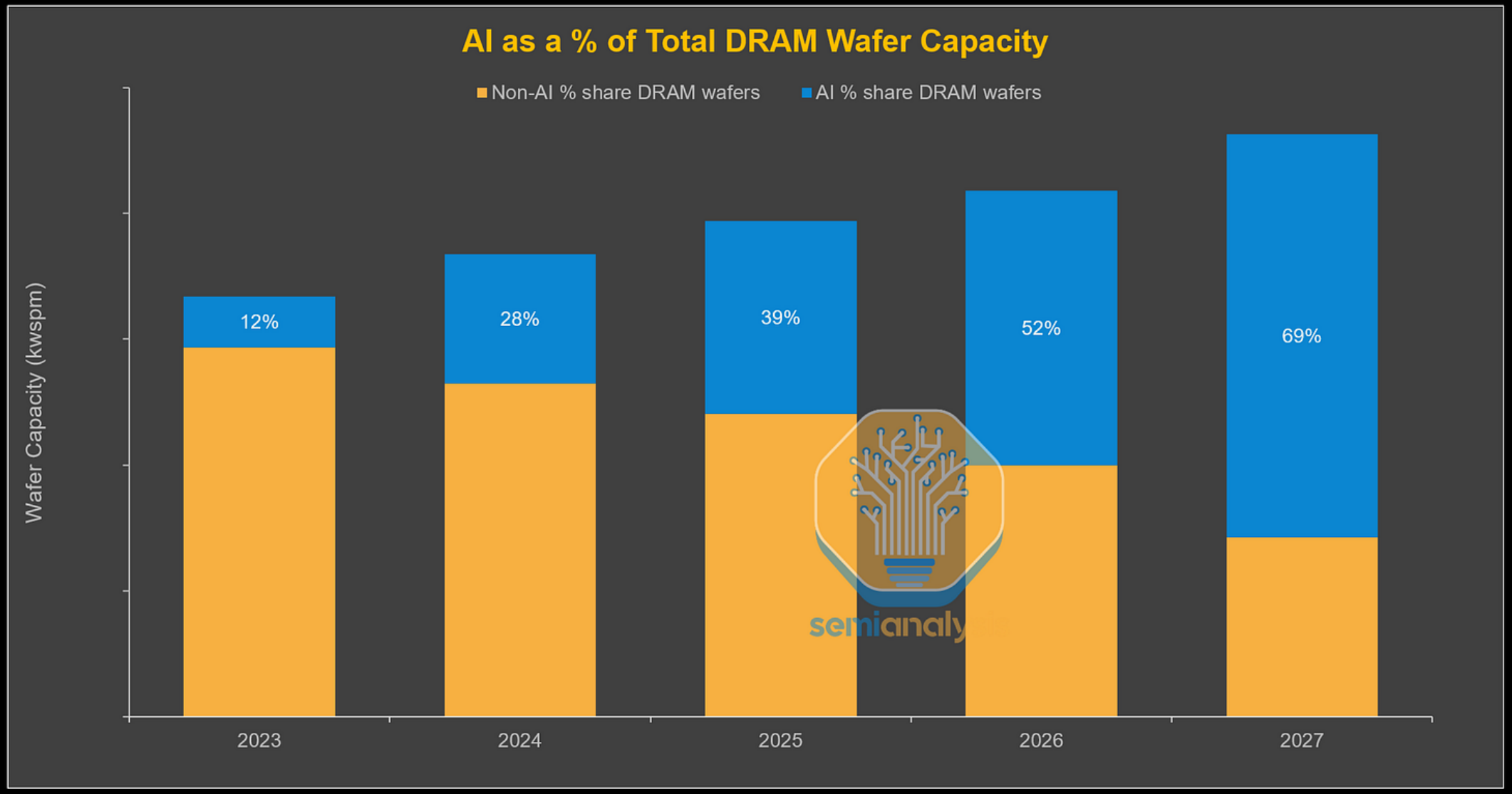
TO BE CLEAR, ALL OF THE GROWTH IS BECAUSE OF AI



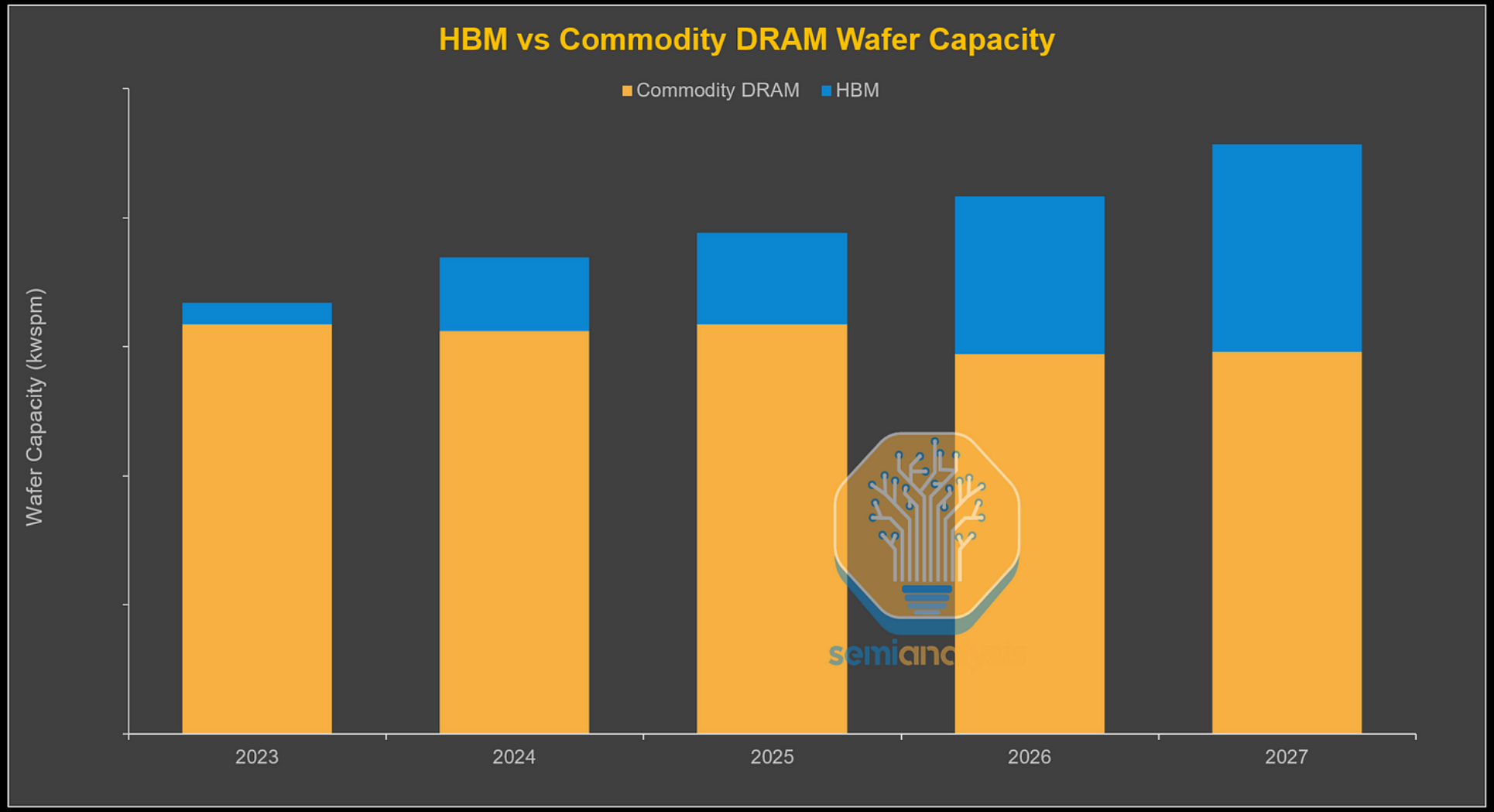
AI GROWTH IS COMING AT THE EXPENSE OF SMARTPHONES



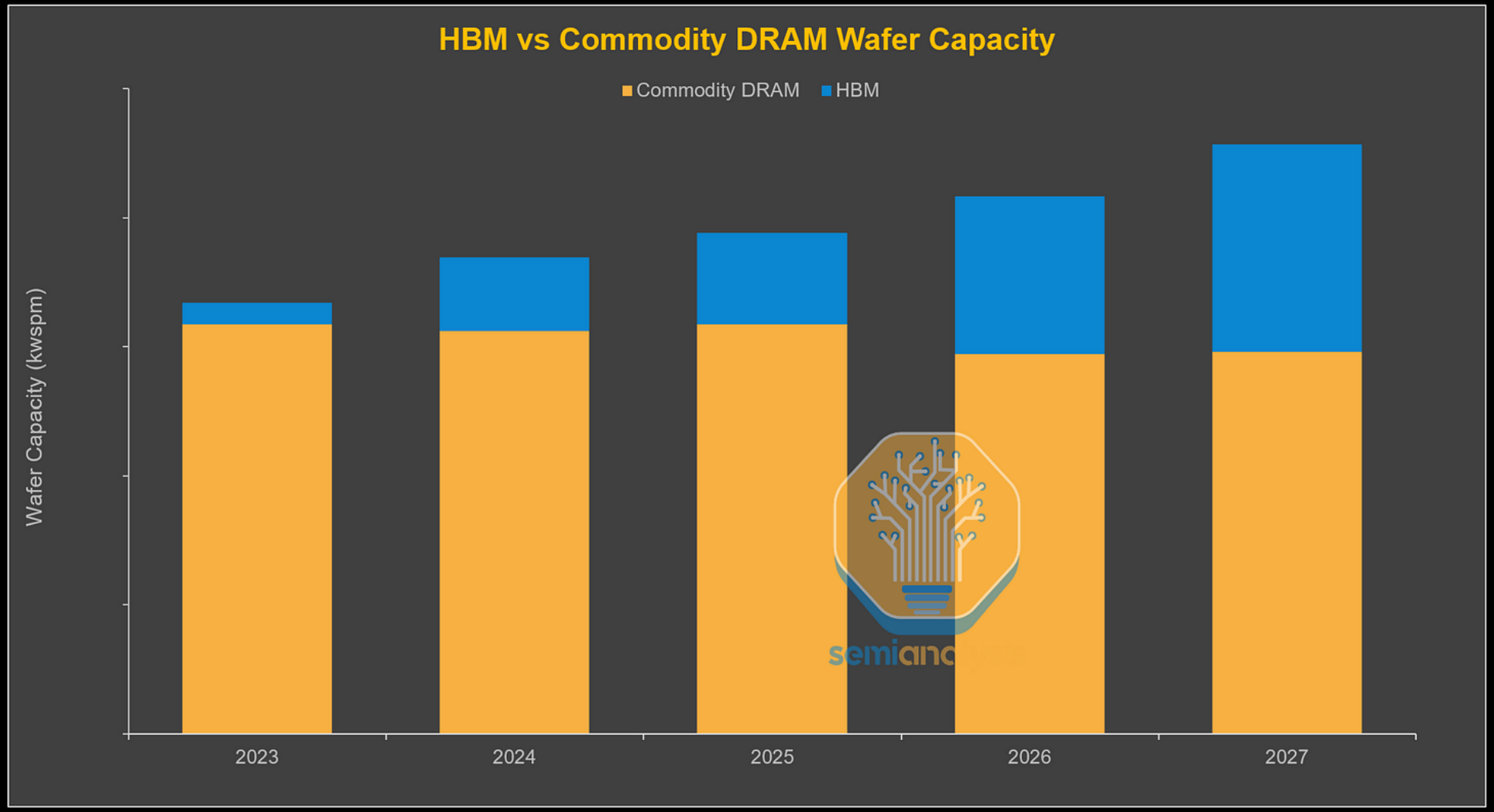
AI IS ALSO COMING FOR MEMORY



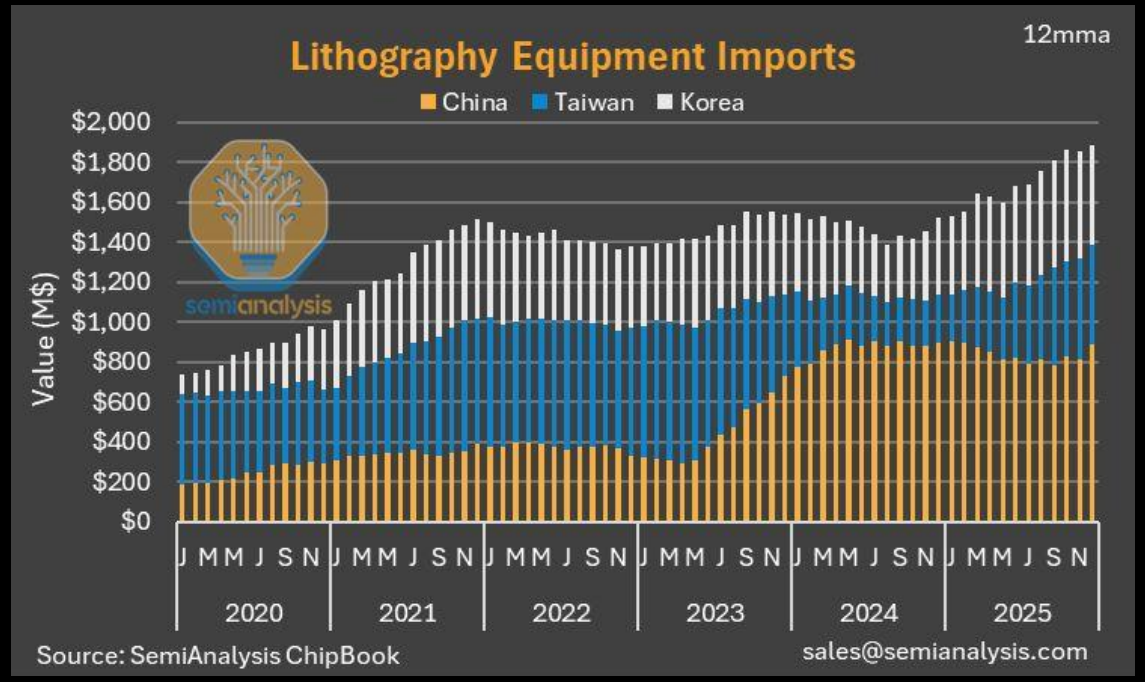
AI IS ALSO COMING FOR MEMORY



AI IS ALSO COMING FOR MEMORY

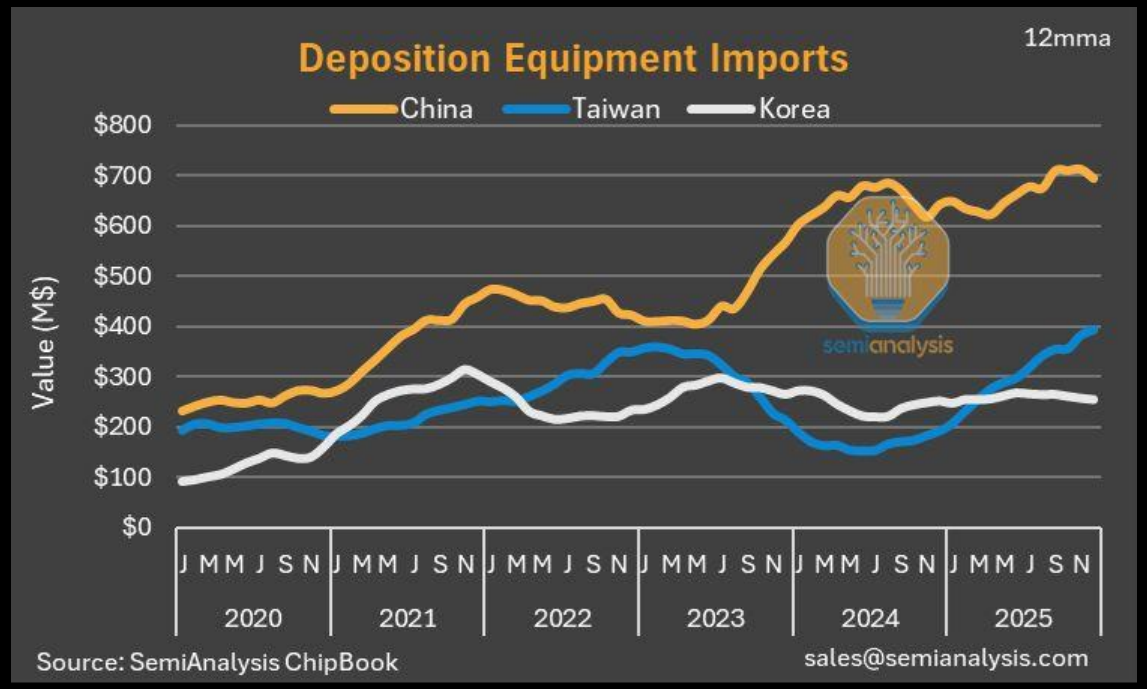


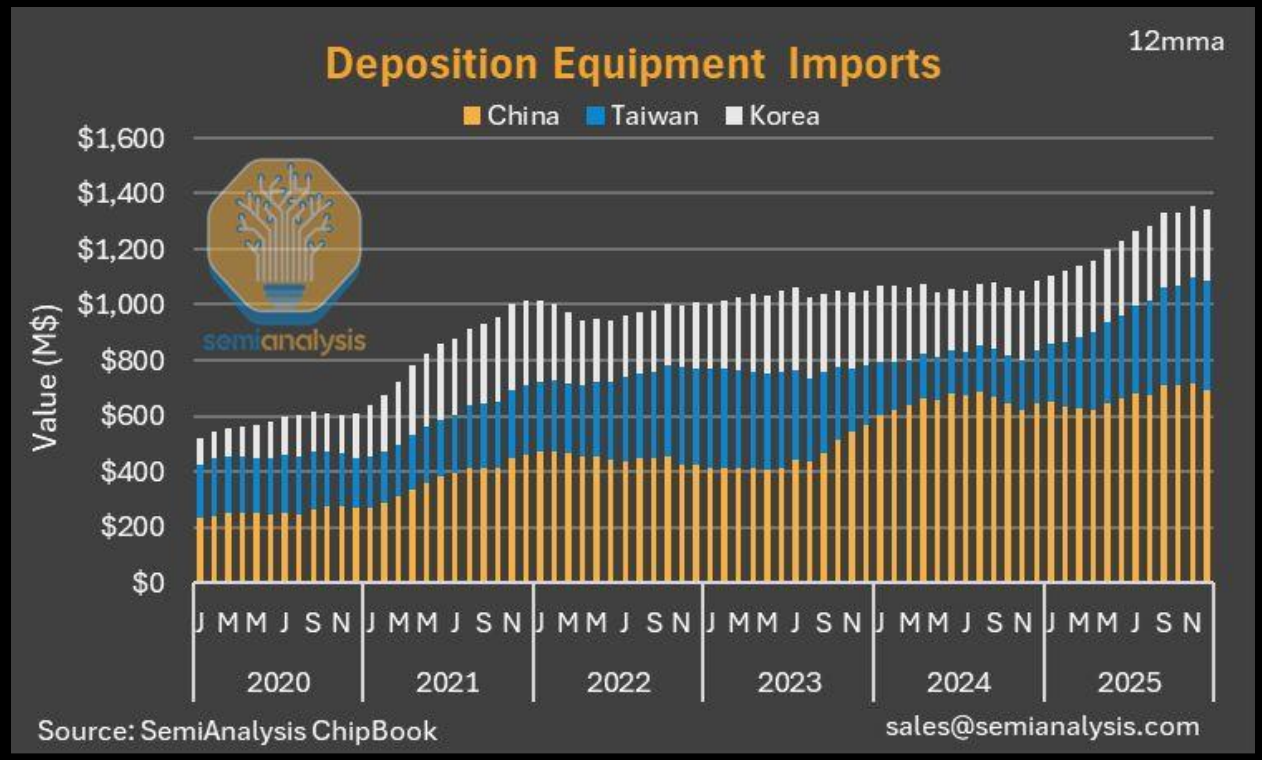
Result: Wafer Fab Equipment (WFE) Spend Is Increasing



Source: SemiAnalysis Wafer Fab Equipment Model and SemiAnalysis Chipbook

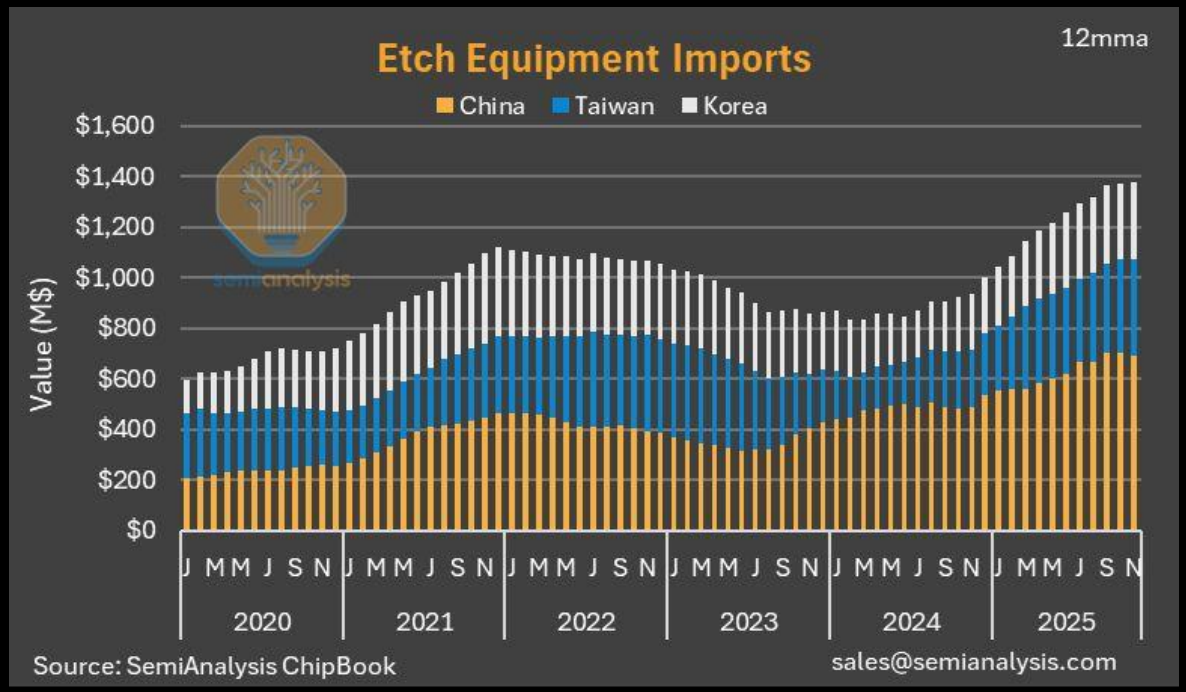
Result: Wafer Fab Equipment (WFE) Spend Is Increasing





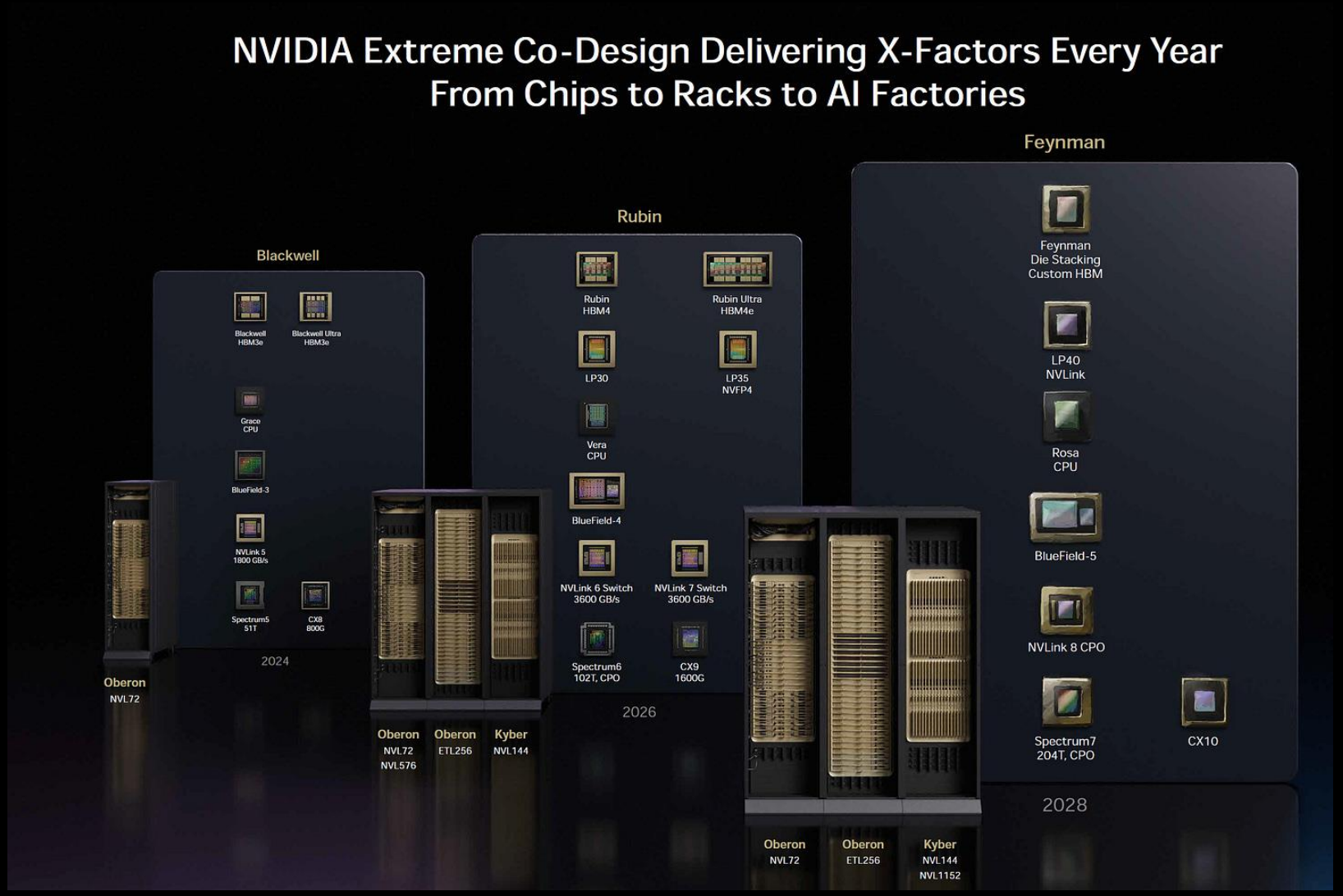
Source: SemiAnalysis Wafer Fab Equipment Model and SemiAnalysis Chipbook

Result: Wafer Fab Equipment (WFE) Spend Is Increasing



Source: SemiAnalysis Wafer Fab Equipment Model and SemiAnalysis Chipbook

NEW GPUS ARE COMING

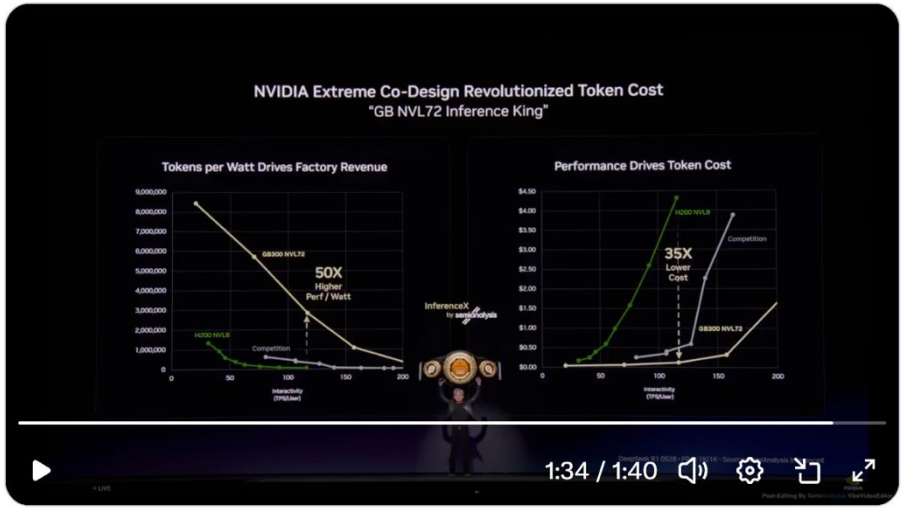


SO FAR, PERFORMANCE IMPROVEMENTS HAVE BEEN IMMENSE

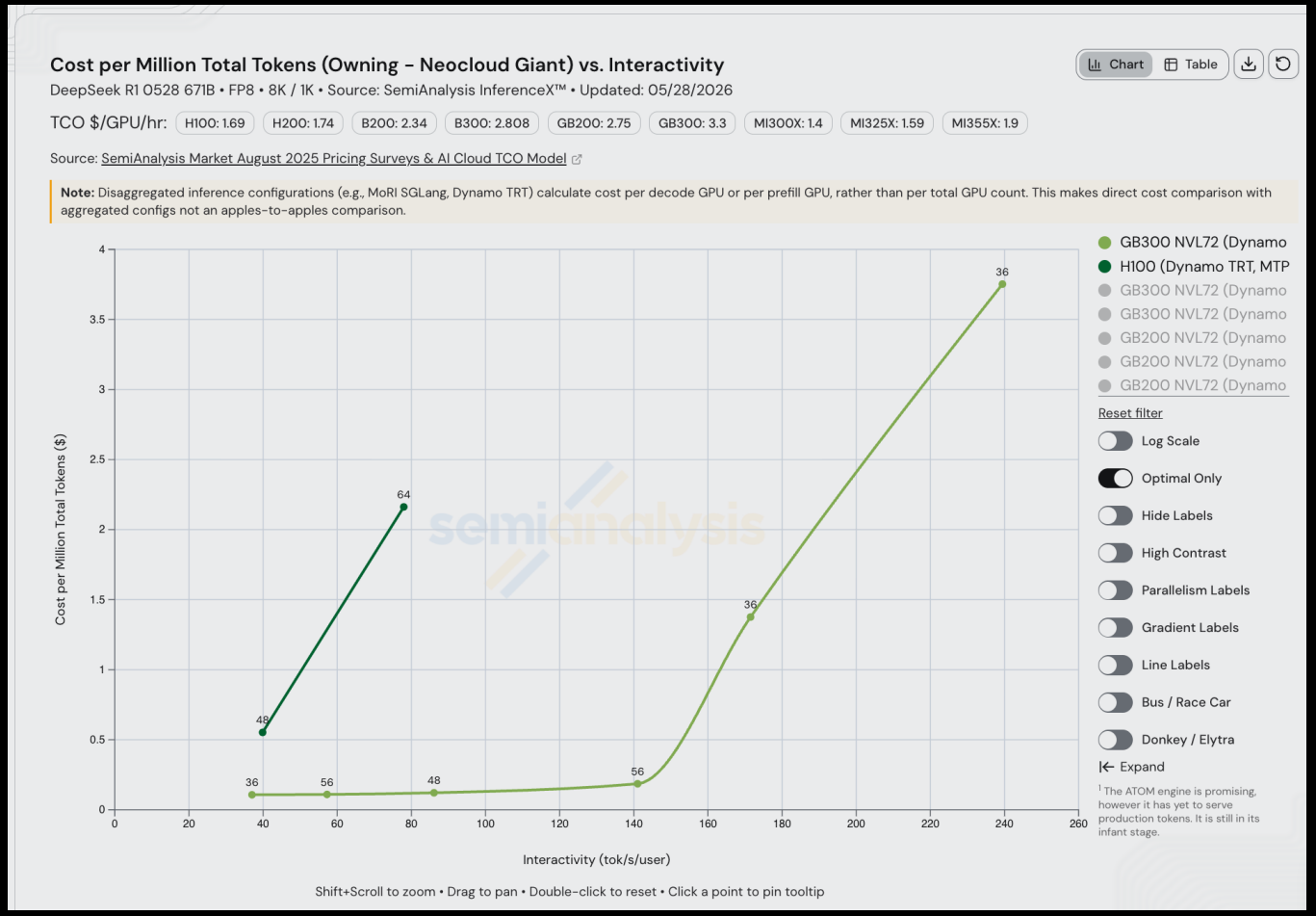
← **Post**

SemiAnalysis ✓
@SemiAnalysis_

At GTC 2024, Jensen said that GB200 NVL72 was 35x faster than Hopper. Nobody believed it and thought it was classic fake Jensen Math. When we tested the performance of it, it wasn't just 35x faster, it was over 50x times faster even against an strong Hopper baseline with all of the inference optimization composed together like MTP, Disagg prefill, wideEP, etc. View the nuanced results at InferenceX dot com.

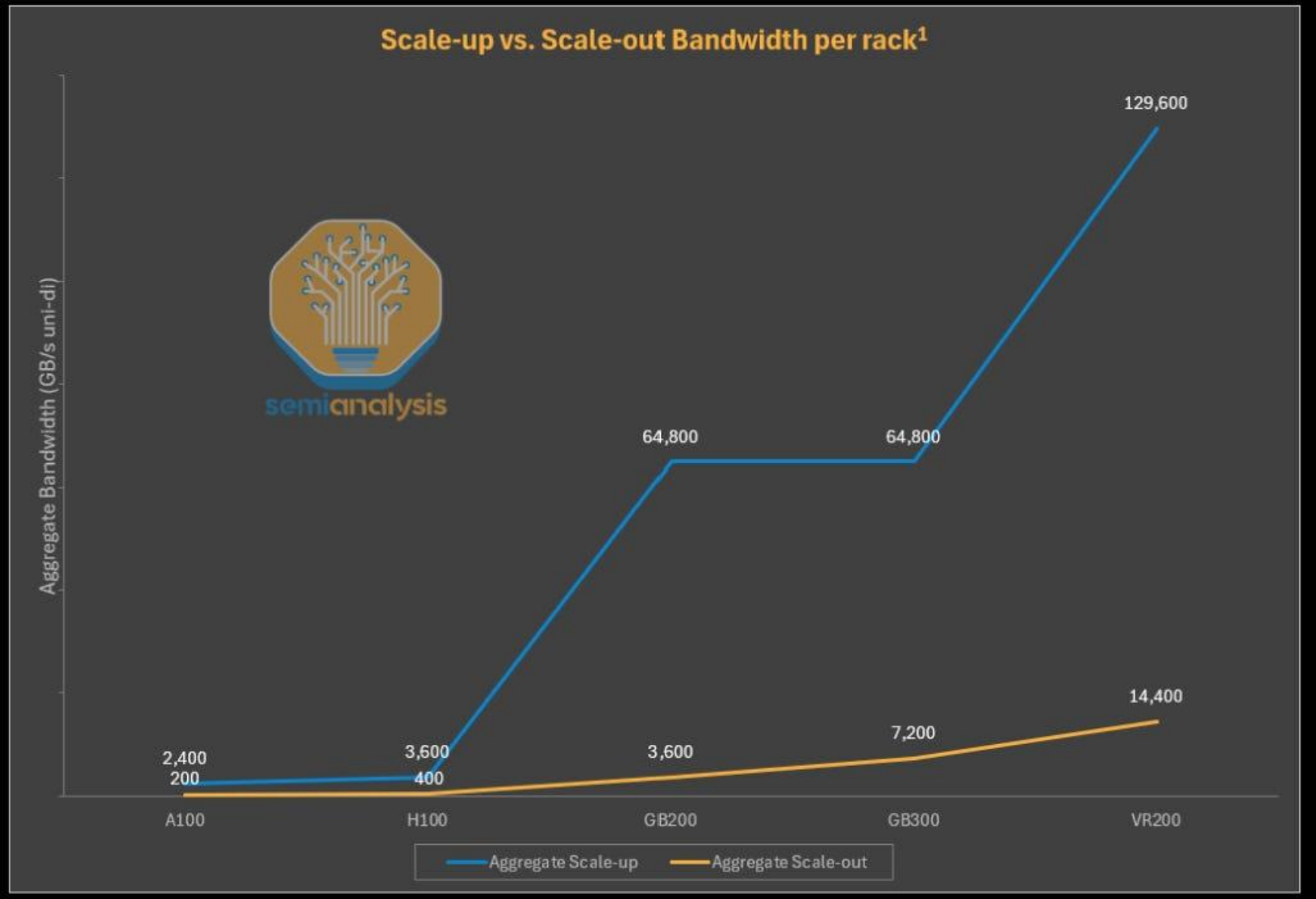


3:07 PM · Apr 18, 2026 · **60.8K** Views



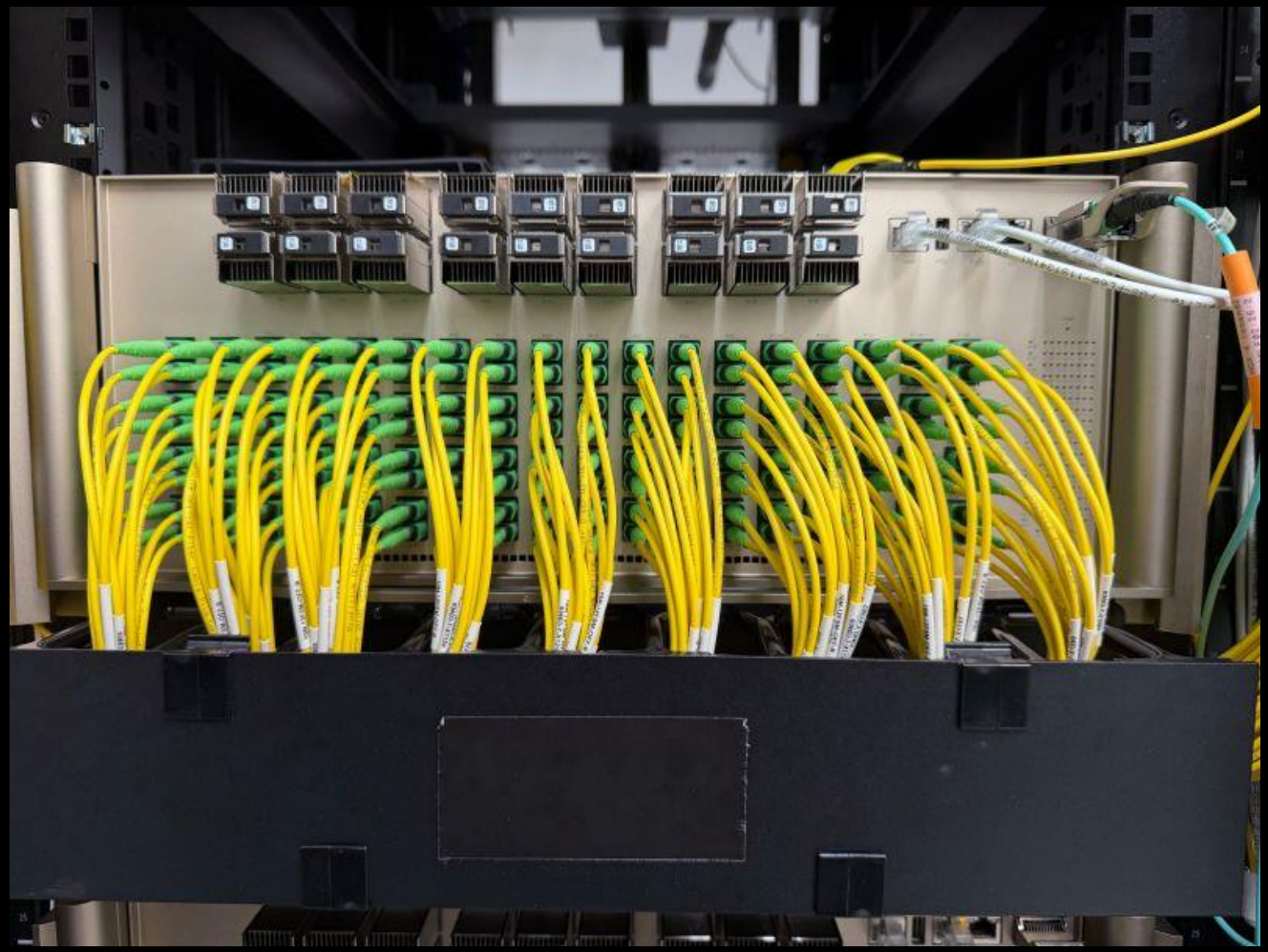
$2/Mtok on H100, $0.10/Mtok on GB300 NVL72 (source: inferencex.com)

NETWORKING IS THE PERFORMANCE BOTTLENECK



Notes: (1) Each rack has 8 GPUs for A100 and H100; each rack has 72 packages for GB200, GB300 and VR200

CPO IS COMING



Source: Lambda

COPPER TODAY, OPTICS TOMORROW

DAC → ACC → AEC → optical — each step trades reach, power, and cost

DAC	ACC	AEC	Optical / Kyber
Passive copper	**Active copper**	**Active electrical cable**	**Co-packaged + pluggable**
Reach	Reach	Reach	Reach
<2 m reach	3–5 m	5–7 m	10–100+ m
Power	Power	Power	Power
0 W per port	1–2 W	3–5 W	8–15 W (LR/FR)
Where	Where	Where	Where
In-rack	Top-of-rack	Adjacent racks · NVL72 backplane	Scale-out · cross-row

NVL72 stays copper. NVL576 / Kyber-class designs need optical reach + co-packaged optics.

Source: SemiAnalysis Networking Model, NVIDIA Kyber announcement

CPO IS A POWER + RELIABILITY BET

Co-packaged optics cuts switch power and reach — but MTTF is the open question

22%

less switch power

CPO removes pluggable DSP + retimers — frees rack thermal budget for more compute.

20%

lower $/Gbps at scale

Higher initial BOM, but lower lifetime energy + cable cost across 100k-port fabrics.

MTTF

is the question

Field MTBF data is still thin; pluggables remain the safer choice for now.

Vendor positioning

NVIDIA	Quantum-X Photonics + Spectrum-X Photonics — first 800G/1.6T CPO switches
Broadcom	Bailly / Tomahawk-6 CPO — pursued, then re-scoped as pluggable-first
Marvell + partners	CPO PHY IP licensing into custom silicon
Status	Pluggable optics still ship the volume — CPO penetrates only at the highest radix tiers

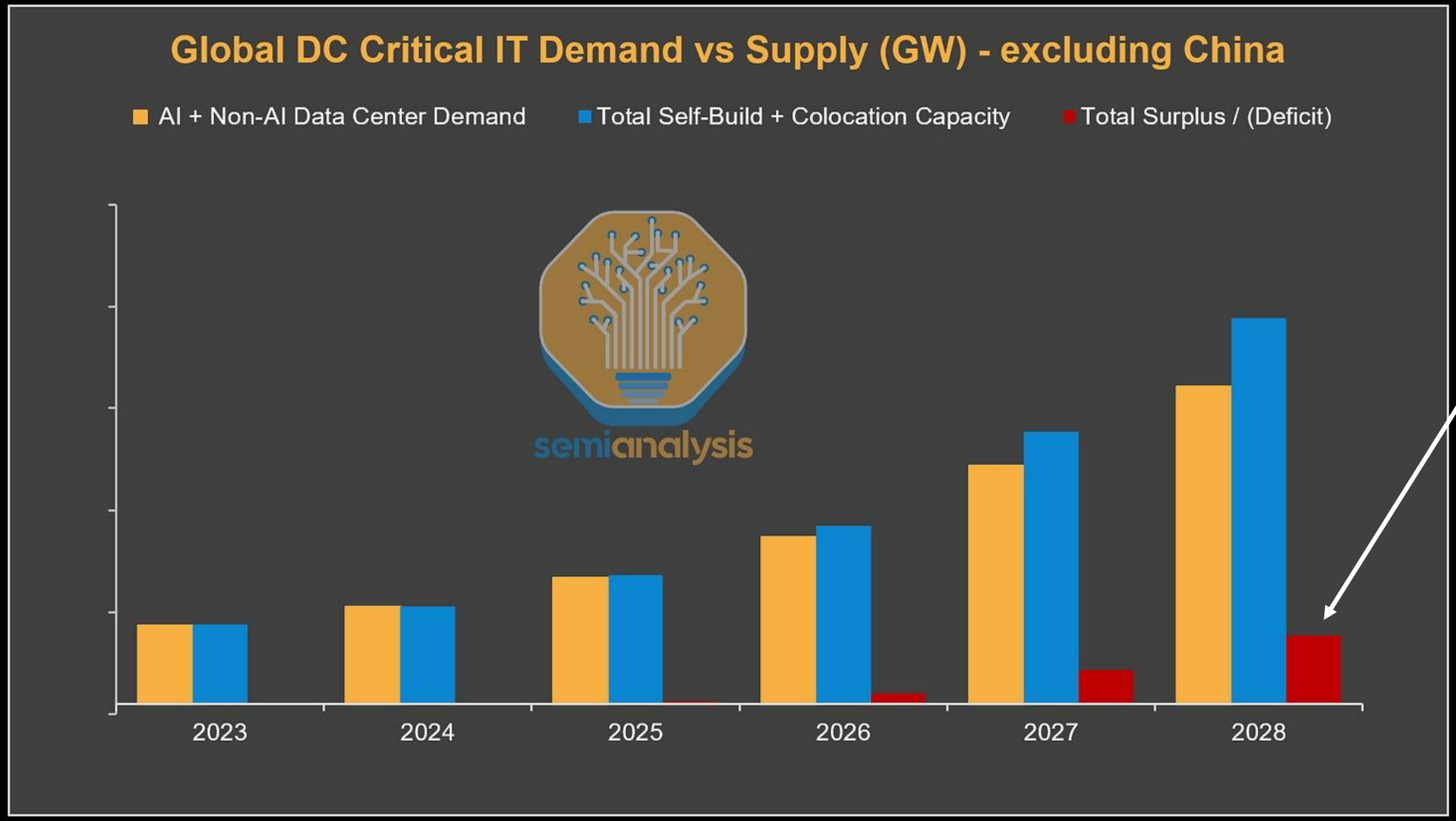
Source: SemiAnalysis Networking Model, NVIDIA Quantum-X / Spectrum-X Photonics

PART 02

Datacenters

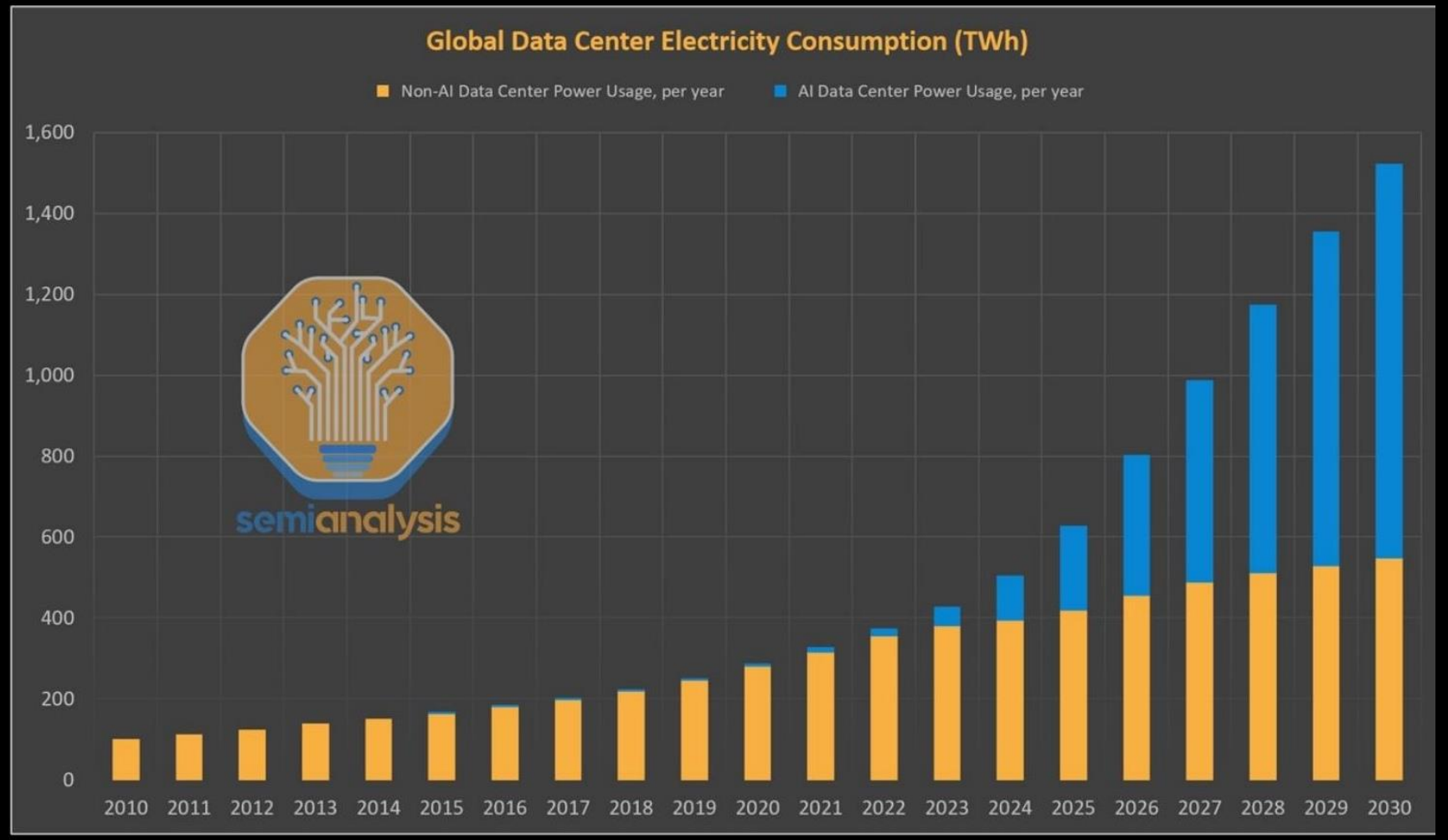
Megawatts per rack, gigawatts per campus — the grid is the new bottleneck.

WHERE DO ALL THESE CHIPS GO?



Expanding datacenters and power is much easier than building new fabs.

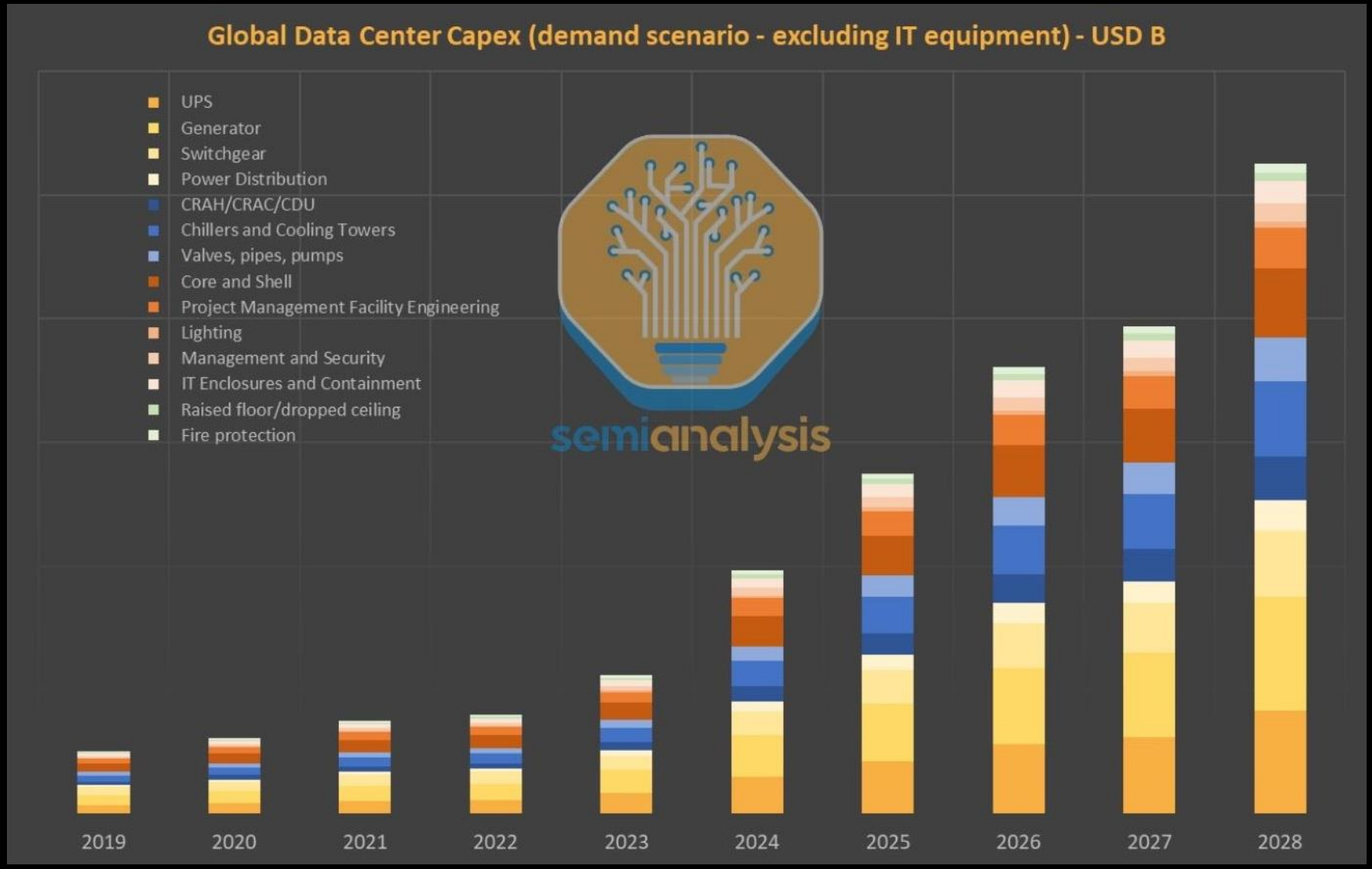
Industry-wide load growth — GW additions by region



Forecast critical IT load additions by region and customer (hyperscaler, neocloud, enterprise) through 2030.

Source: SemiAnalysis Datacenter Model

Industry-wide load growth — GW additions by region



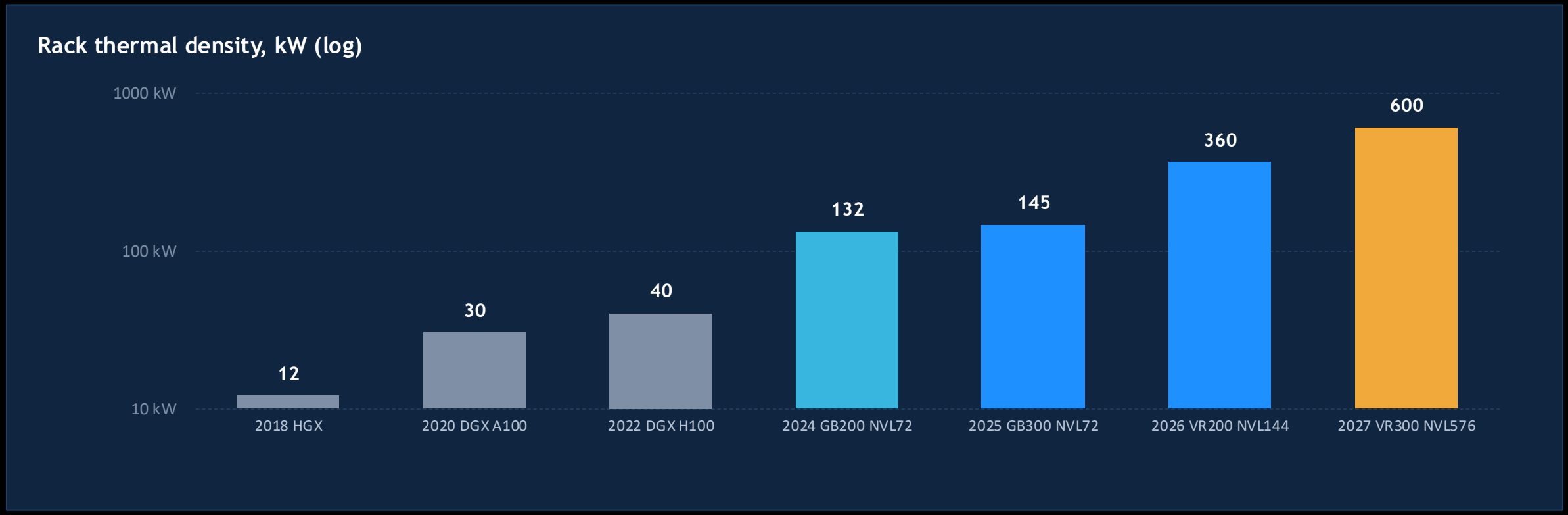
Source: SemiAnalysis Datacenter Model

YOU CAN GO SEE IT FOR YOURSELF

(pictures from Friday last week)

RACK POWER HAS GONE EXPONENTIAL

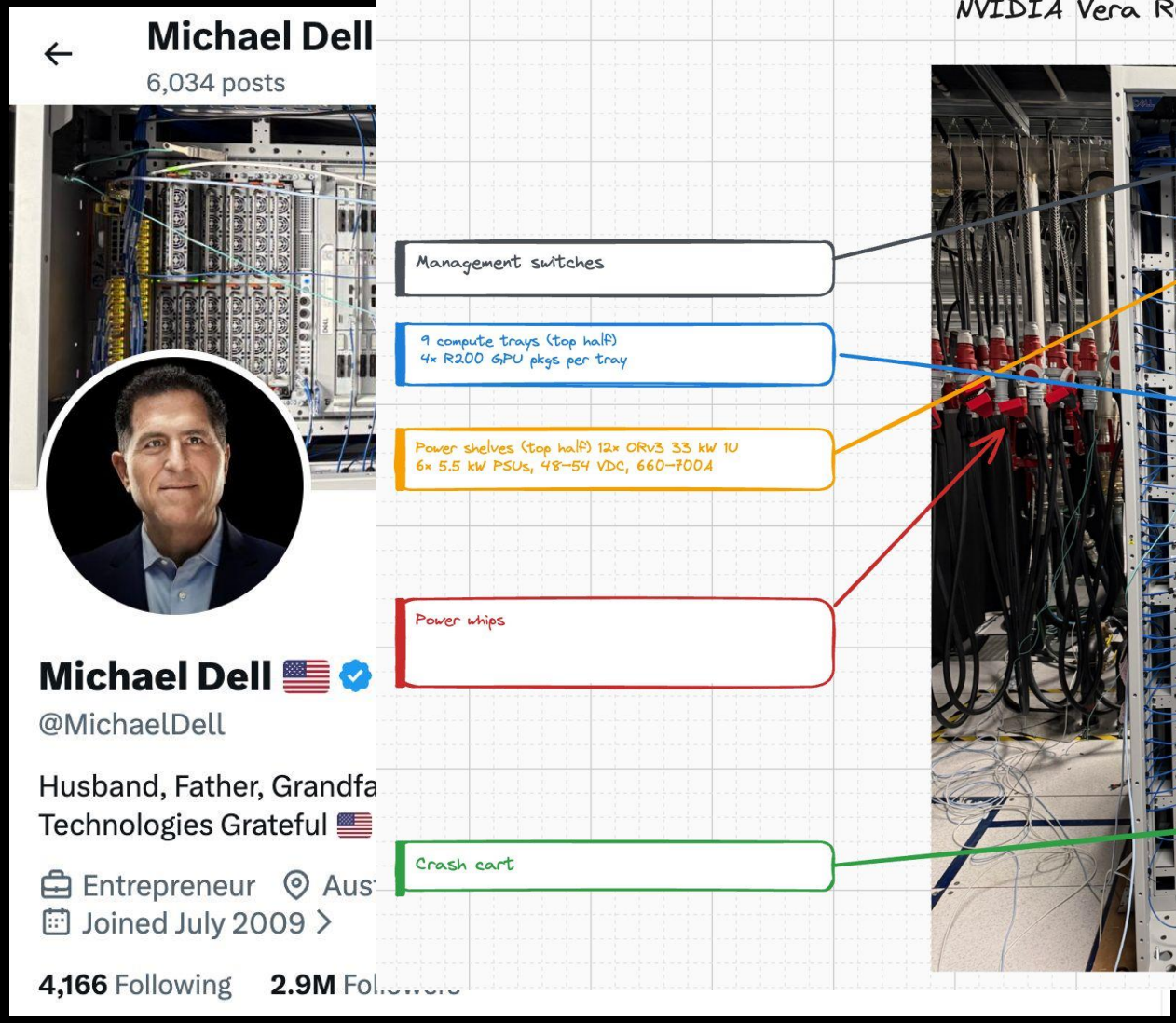
From 12 kW air-cooled to 600+ kW liquid-cooled in one product generation



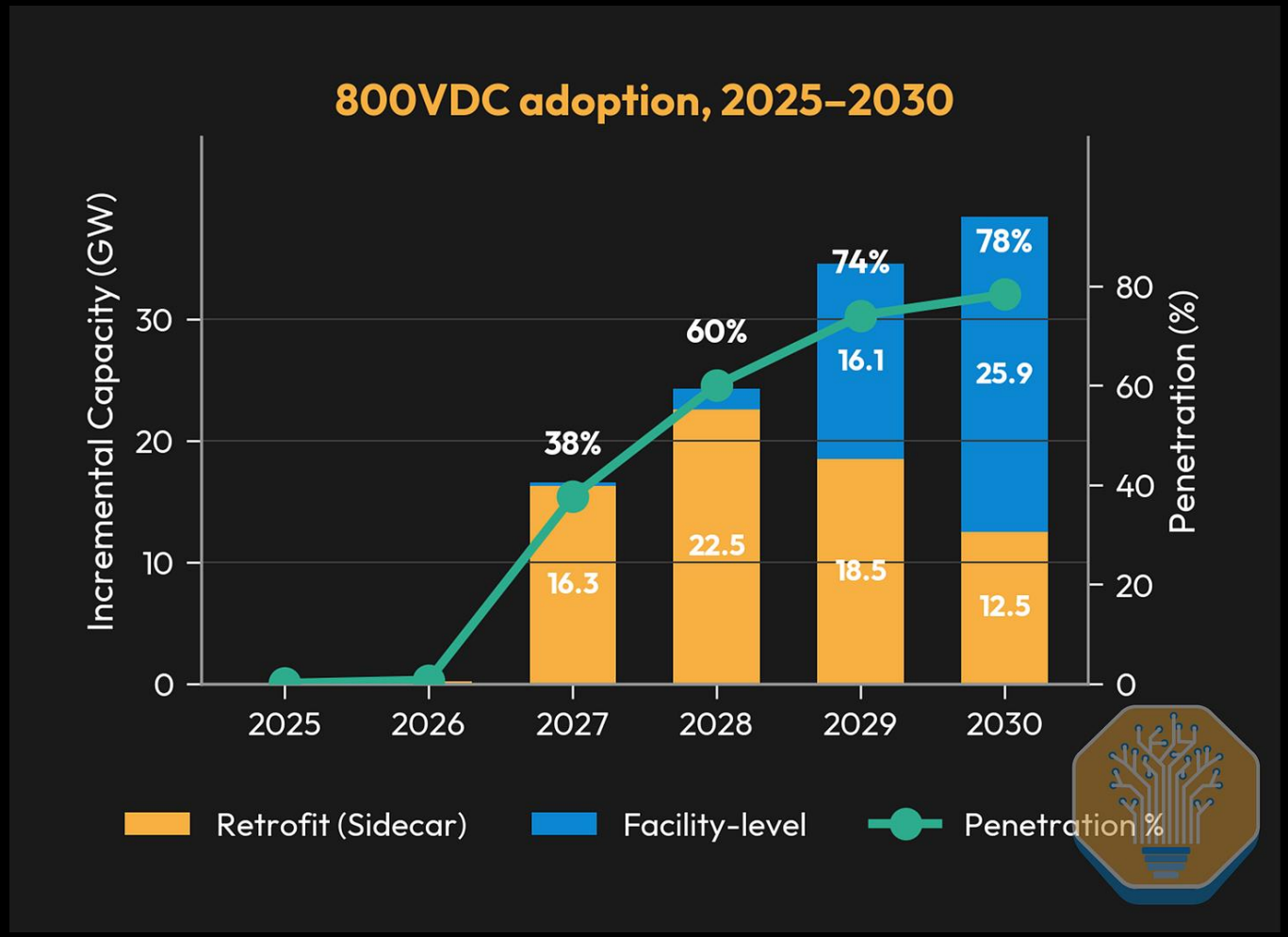
50× density growth in 9 years — air cooling cannot follow this curve.

Source: NVIDIA datasheets, SemiAnalysis Datacenter Model

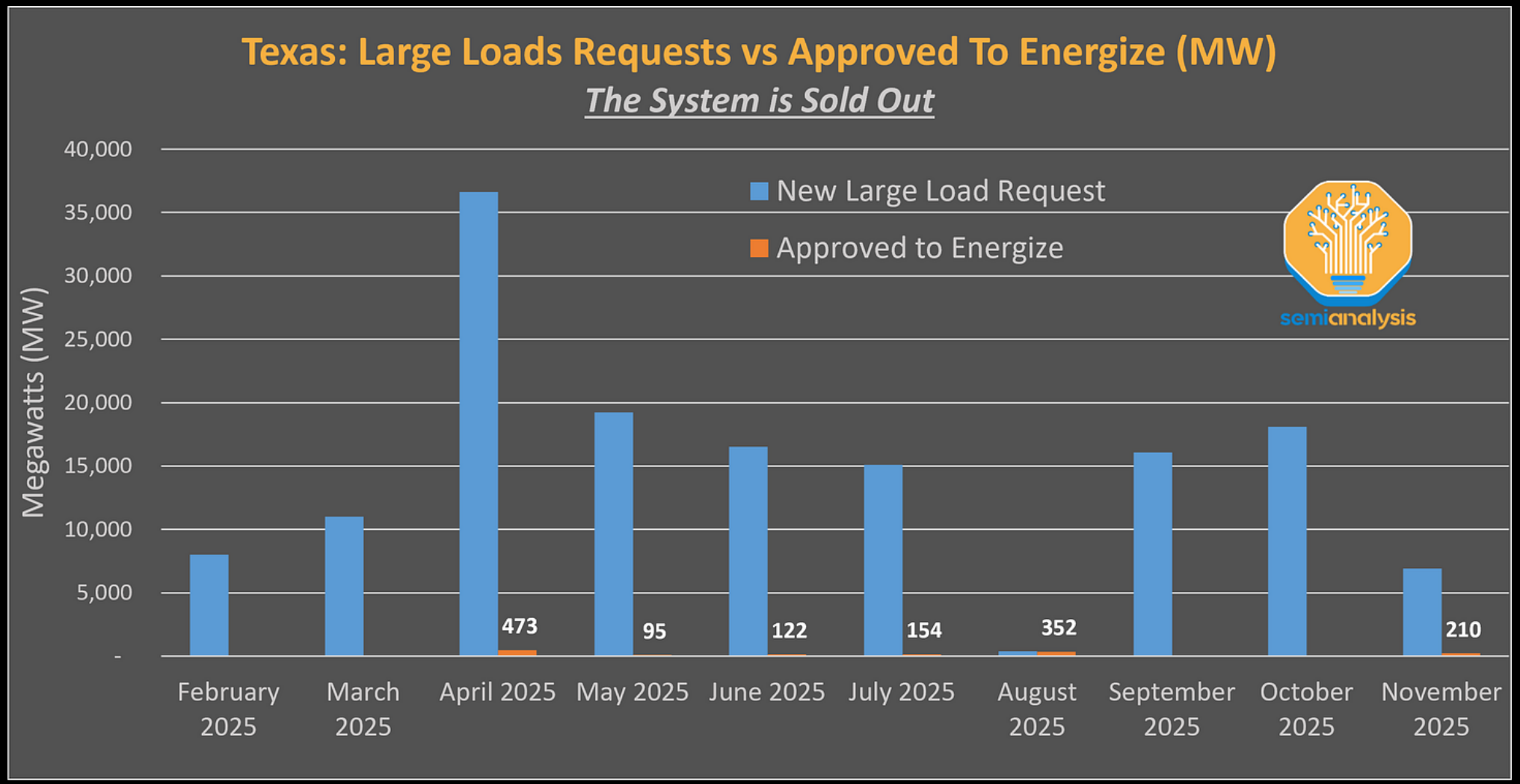
THESE GPUS ARE COMING



THE RESULT IS 800VDC



BUT THE GRID IS SOLD OUT



THE SOLUTION: BEHIND-THE-METER GAS

xAI Colossus 1 - twelve SMT-130 turbines (198MW)
Source: SemiAnalysis Datacenter Industry Model

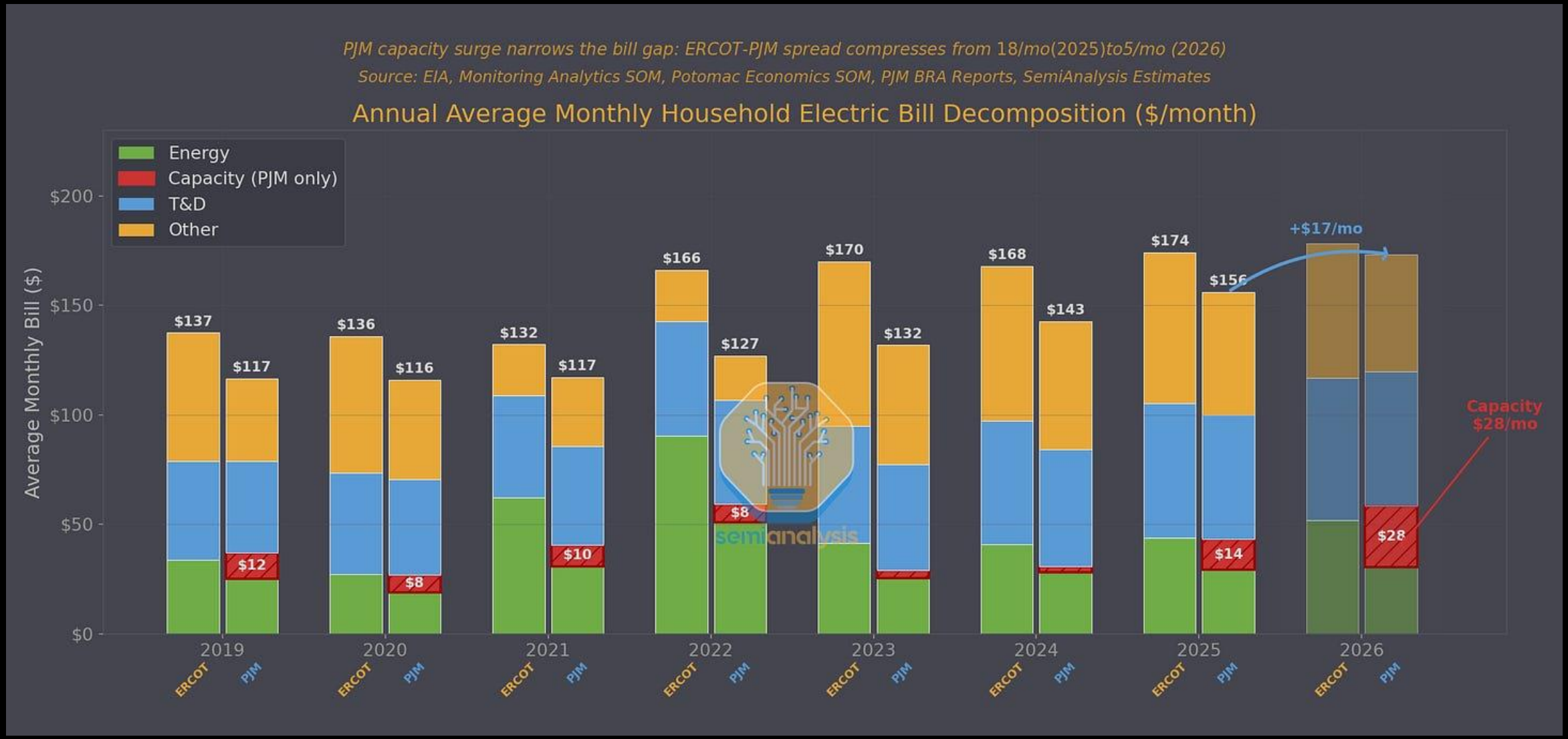
xAI Colossus 2 - seven Titan 350 turbines (266 MW)
Source: SemiAnalysis Datacenter Industry Model

Construction tracking — satellite + permits

Satellite imagery + permit + interconnect data gives us a real-time view of construction progress across every major AI campus.

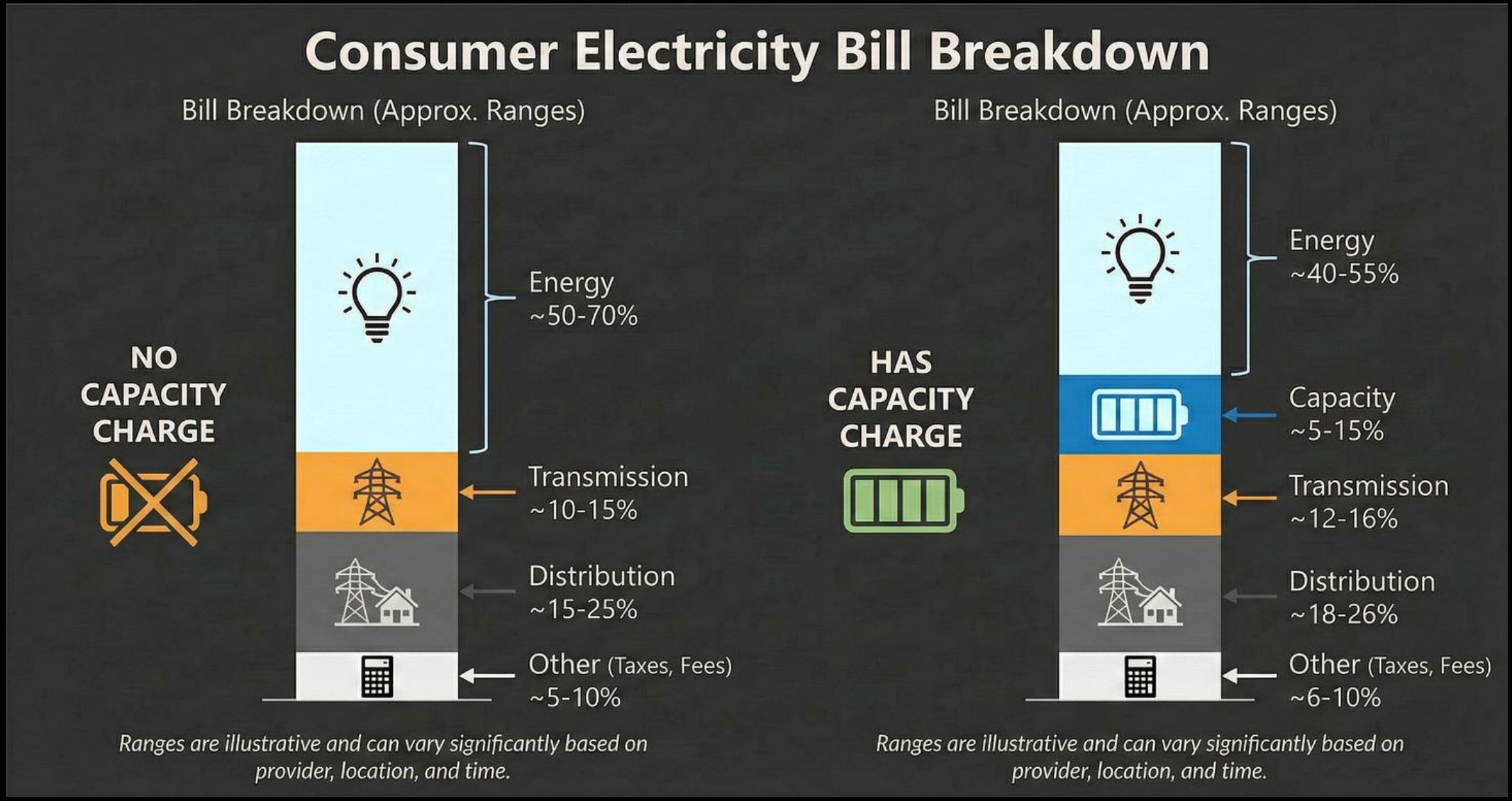
Source: SemiAnalysis Datacenter Model

Are AI Datacenters Increasing Electric Bills for American Households?



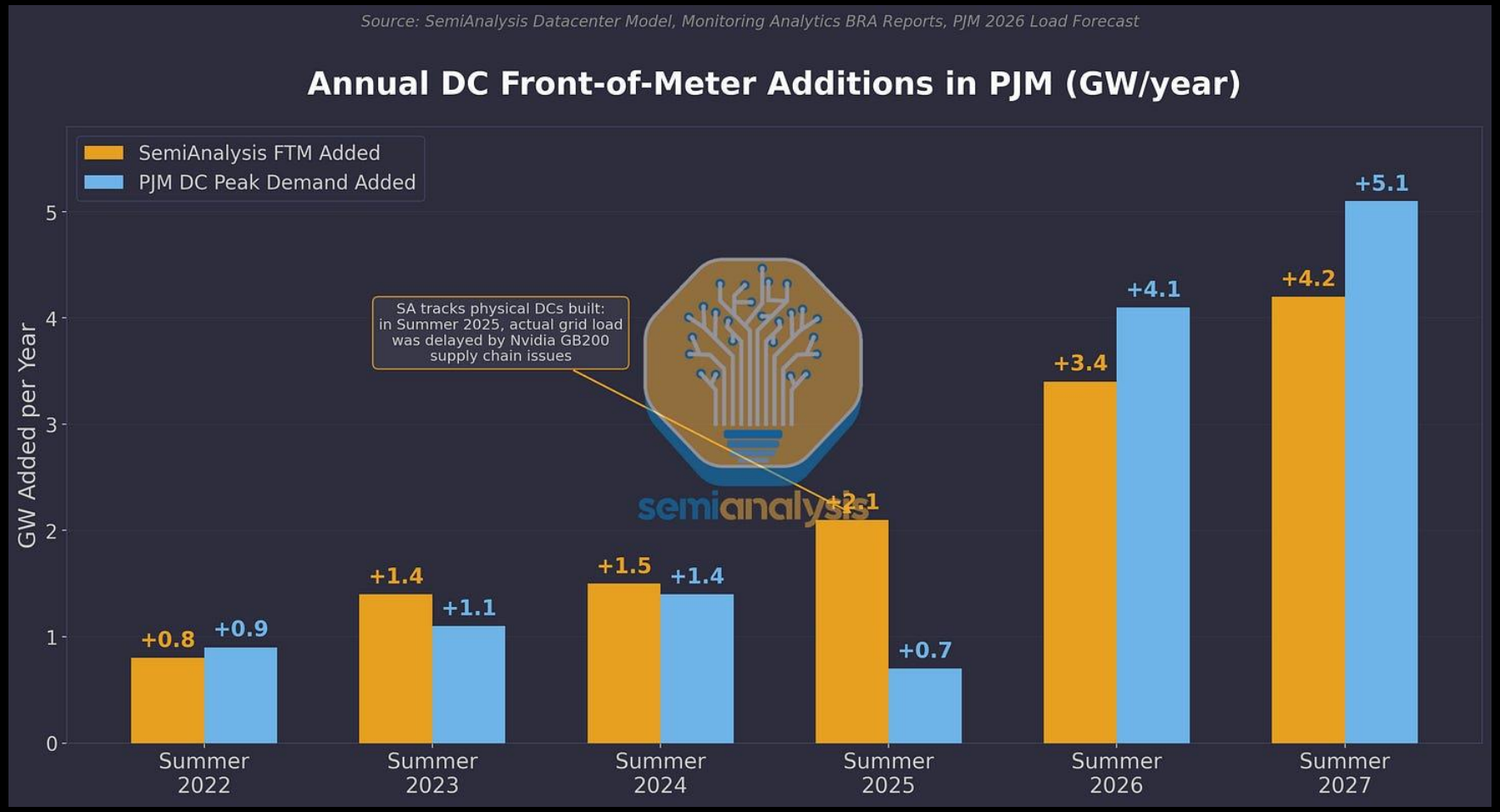
Source: SemiAnalysis Energy Model

Are AI Datacenters Increasing Electric Bills for American Households?



Source: SemiAnalysis Energy Model

Energy Model — generation mix, interconnect, behind-the-meter

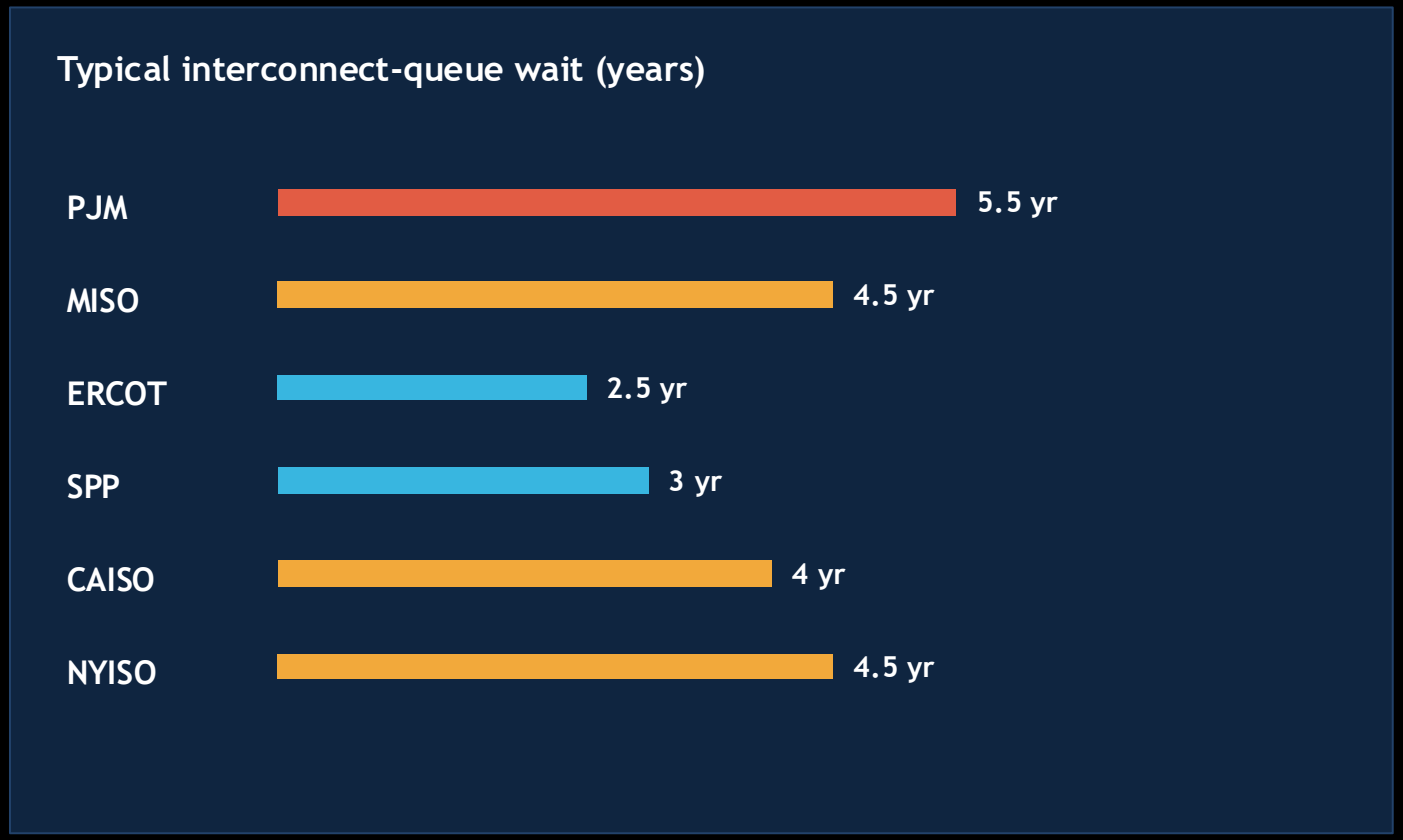


Power generation, transmission, and behind-the-meter capacity additions feeding AI data center demand through 2030.

Source: SemiAnalysis Energy Model

BEHIND-THE-METER IS NOW THE PATH

Grid interconnect queues stretch 4–7 years — speed-to-energy decides who trains first



Who is going behind-the-meter	
xAI Memphis	Gas turbines on-site, ~150 MW added in months
OpenAI Stargate	Multi-GW build with co-located generation
Meta Hyperion	Louisiana, gas + storage hybrid
Crusoe + OAI TX	Pad-side gas — fast permit path
Anthropic + AWS	Nuclear PPAs for baseload, multi-year horizon

Source: SemiAnalysis Energy Model, FERC Order 2023

PART 03

System Performance

Two independent, public projects: ClusterMAX compares neoclouds, InferenceX compares chips

WHAT IS A NEOCLOUD?

Forbes

INNOVATION > CLOUD

Neoclouds' Rise Reflects How AI Is Tra
The Cloud Market

By Matt Kimball, Contributor. Senior Analyst, Servers for Moor Insights and Strategy

Follow Author

Published Mar 25, 2026, 12:49pm EDT

EQUINIX | Interconnections

THE INFRASTRUCTURE BEHIND AI

What Is a Neocloud?

Forbes

INNOVATION > CLOUD

Inside The Neocloud Economy: What's Next For GPU-As-A-Service

McKinsey & Company

Tech & AI

How We Help Clients | Our Insights | Case Studies | Contact Us

The evolution of neoclouds and their nex moves

November 19, 2025 | Article

By Massimo Mazza, Pankaj Sachdeva, Suren Arutyunyan, and Tarik Alatovic

Share Print

Born out of GPU scarcity, neoclouds now face a harder test: E
beyond bare metal or risk repeating Cloud 1.0 history.

McKinsey & Company – Listen to the article: The evolution of neoclouds and their next moves

Home > Artificial intelligence, automation and robotics

OPINION

Gartner: Why neoclouds are the future of GPU-as-a-Service

Neoclouds are set to change the economcs of AI in the cloud. By 2030, neocloud providers will capture around 20% of the $267bn AI cloud market

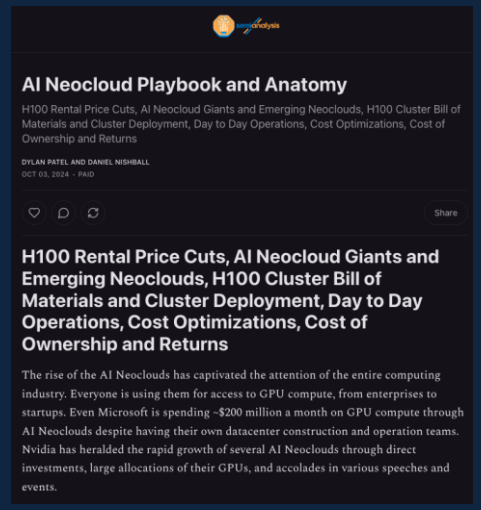
By **Mike Dorosh** Published: **20 Feb 2026**

For the past decade, hyperscalers have defined how CIOs and IT leaders think about their organisation's cloud infrastructure. Scale, abstraction and convenience became the default answers to almost every compute question. But artificial intelligence (AI) is breaking the economics of cloud computing and neoclouds are emerging as the response.

Gartner estimates that by 2030, neocloud providers will capture around 20% of the $267bn AI cloud market. Neoclouds are purpose-built cloud providers designed for graphics

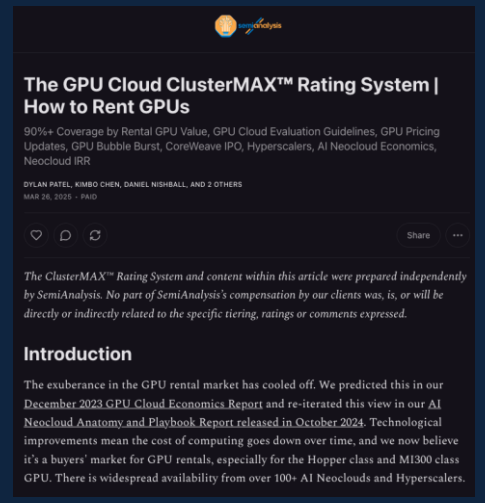
semianalysis

HOW WE GOT HERE



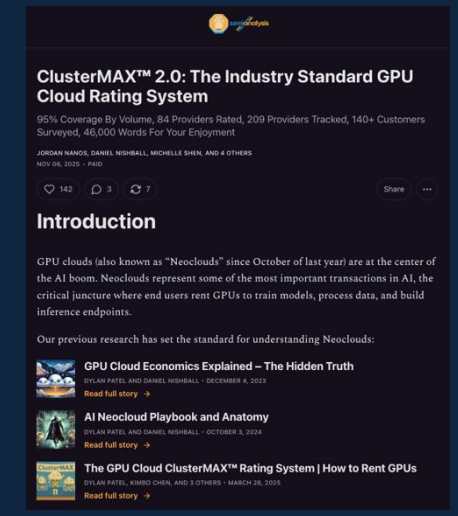
SemiAnalysis coins the term “Neocloud”

October 3, 2024



The first ClusterMAX rating system is released

March 26, 2025



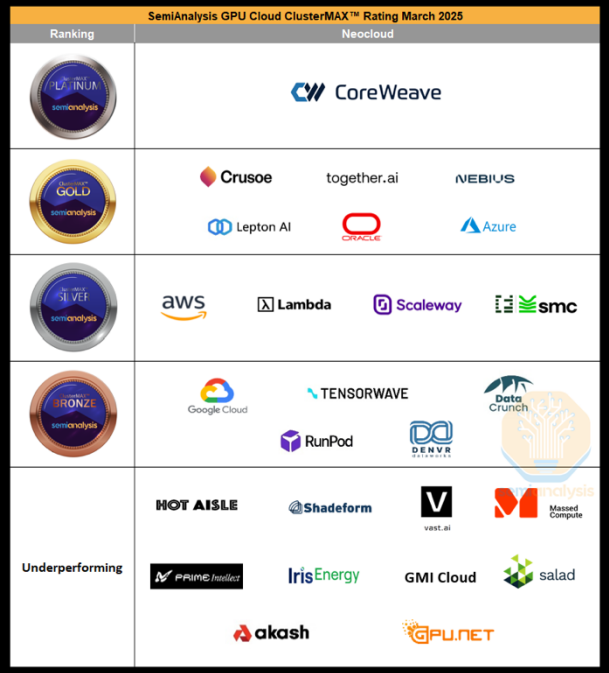
ClusterMAX 2.0 is released

November 6, 2025

CURRENT RANKINGS

ClusterMAX 1.0

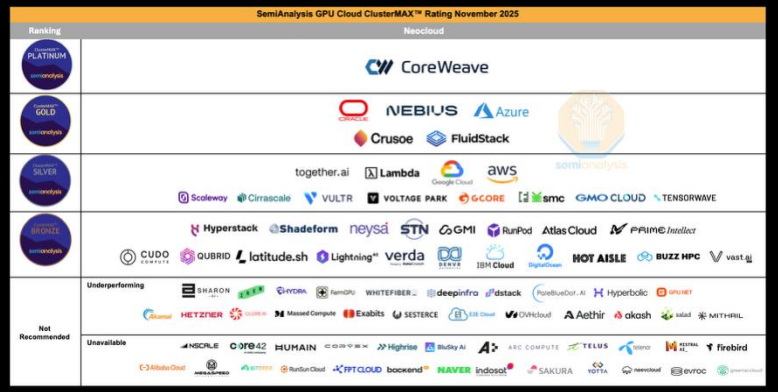
(March, 2025)



26 providers rated

ClusterMAX 2.0

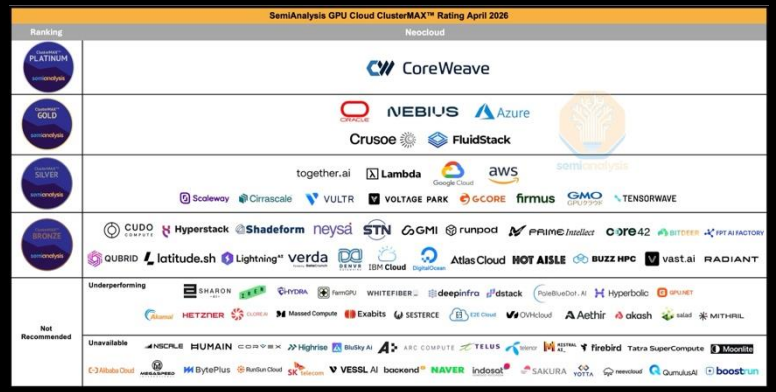
(November, 2025)



84 providers rated

ClusterMAX 2.1

(April, 2026)



90 providers rated

WHAT MAKES A GOOD NEOCLOUD?

Expectations

Explore our detailed expectations for different deployment models and infrastructure components

- SLURM
- Kubernetes
- Standalone
- Monitoring
- Health Checks

Criteria

Explore the 10 critical dimensions we use to evaluate GPU cloud providers

- Security
- Lifecycle
- Orchestration
- Storage
- Networking
- Reliability
- Monitoring
- Pricing
- Partnerships
- Availability

Source: https://www.clustermax.ai/

EXAMPLE TESTS WE RUN

Cluster audit, review of monitoring dashboard

Common issues for slurm:

- passwordless ssh to worker nodes, sudo access
- nvidia container toolkit installed and up to date
- ncu installed with hardware counters accessible
- high bw NICs/HCAs are named correctly and not conflicting, allowing NCCL to pick them up: mlx5_0, mlx5_1, etc.
- cuda, nvcc, hpc-x, nccl pre-installed from hpc sdk
- lmod installed and configured
- pyxis, enroot, docker installed and up to date
- topology.conf properly configured
- dcgm's background health checks enabled and plugged into slurm's HealthCheckProgram
- prolog and epilog scripts are lightweight, takes less than 30 seconds to get a node
- dashboard includes slurm job accounting data via sacct for resource usage, summaries by user and group

Common issues for kubernetes:

- passwordless ssh to worker nodes, sudo access
- nvidia container toolkit installed and up to date via GPUOperator or equivalent
- ncu installed with hardware counters accessible
- high bw NICs/HCAs are named correctly and not conflicting, allowing NetworkOperator to work with rdma: 1 pod spec (or equivalent)
- kubeconfig is a simple download
- helm access available without custom external authentication
- csi provider has ReadWriteMany volume support
- default StorageClass configured and functional, PVCs provision without hanging on helm install

EXAMPLE TESTS WE RUN (CONT.)

Real-world workloads

- torchtitan pretraining example: https://github.com/pytorch/torchtitan
- prime-rl example: https://github.com/PrimeIntellect-ai/prime-rl
- inferencex example: https://github.com/SemiAnalysisAI/inferencex

- Networking
  - nccl-tests: https://github.com/NVIDIA/nccl-tests (or rccl-tests) allreduce, allgather, alltoall, etc.
  - ib_write_bw and ib_read_bw from perftest: https://github.com/linux-rdma/perftest
  - stas00 all_reduce_bench.py (similar to nccl-tests, but launched from torch without mpi)

- Storage
  - fio: https://github.com/axboe/fio sequential read/write and random read/write IOPS
  - custom torch.save and dcp.save test
  - custom synthetic data generation test
  - time to run “import torch” on home directory

EXAMPLE TESTS WE RUN (CONT.)

Compute

- nvbandwidth: https://github.com/NVIDIA/nvbandwidth h2d, d2h, p2p
- stas00 mamf-finder.py custom GEMM benchmark from torch, triton, deepgemm

Lifecycle

- synthetic download/upload speed (speedtest)
- hf download/upload (model files)
- container pull/push (ngc, ghcr, dockerhub)
- (uv) pip install torch
- add new users

Reliability

- 8hr burn-in checking for thermal, power, ECCs, link flaps, storage mounts, etc.
- Xid error detection
- sudo reboot

CLUSTERMAX RATINGS GO BEYOND HANDS ON TESTING

To accurately assess providers on reliability, performance, support and pricing at scale we interview hundreds of users

We also introduced a methodology for measuring the impact of downtime, aka “goodput” in our latest article:

Modern GPUs are unbelievably expensive. A single Blackwell GPU costs more than the average car, and uses more energy than a single family home. It is now common for unicorn startups to have thousands of these GPUs working for them, day and night. Many foundation model companies now spend an order of magnitude more money on GPUs than they do on employees. We know multiple companies spending over 80% of their initial funding on GPUs. Startup founders now have four important categories of spending to consider when building a financial plan for their company:

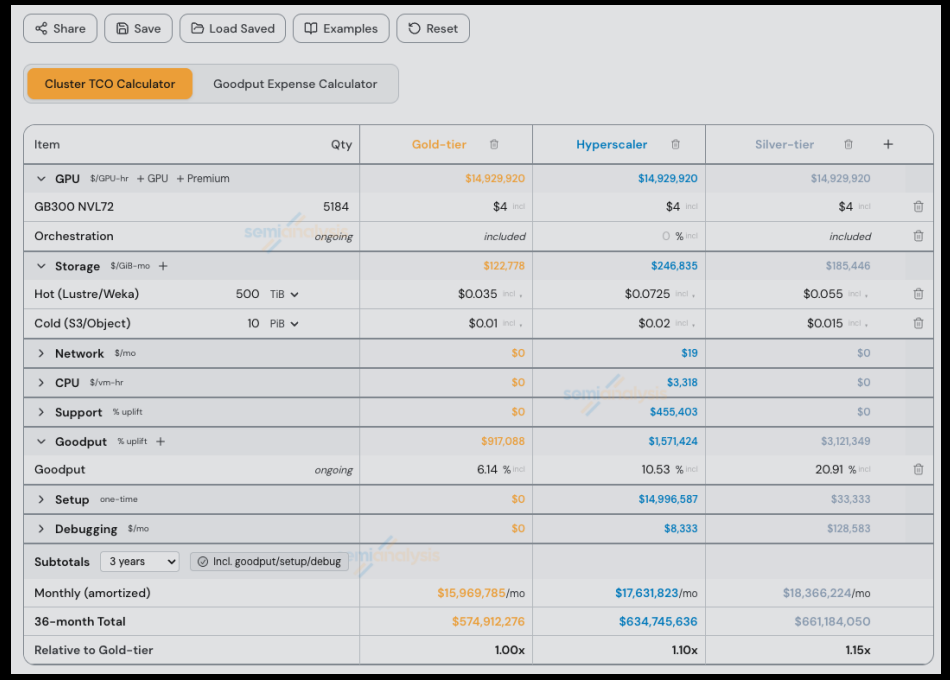
1. GPU clusters
2. Tokens
3. Employees
4. Everything else

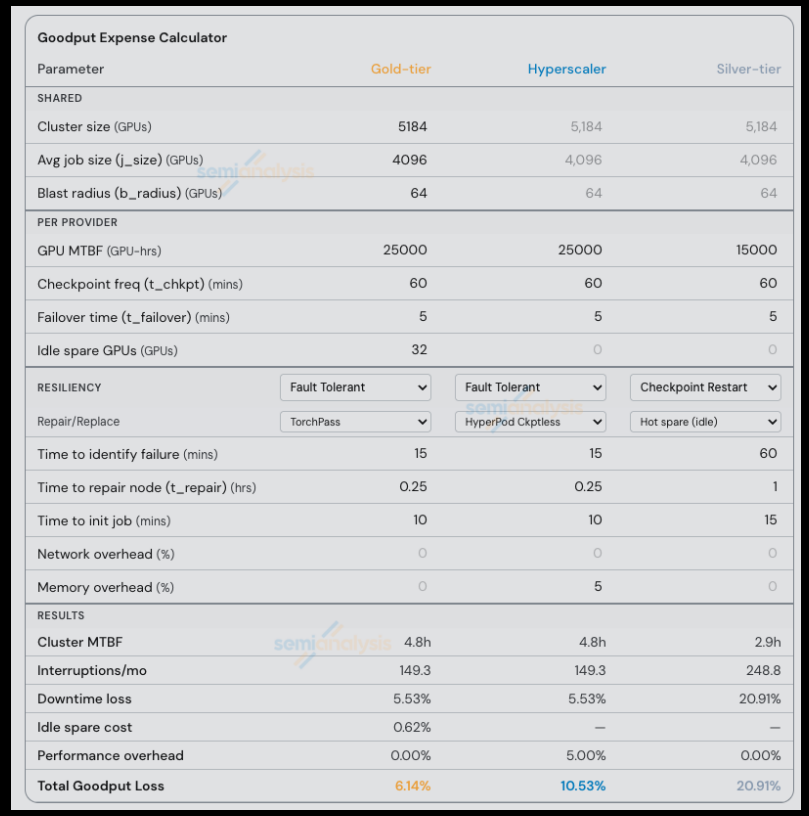
Source: https://clustermax.ai/tco

CLUSTERMAX RATINGS GO BEYOND HANDS ON TESTING

To accurately assess providers on reliability, performance, support and pricing at scale we interview hundreds of users

We also introduced a methodology for measuring the impact of downtime, aka “goodput” in our latest article:





Goodput Expense Calculator

Parameter	Gold-tier	Hyperscaler	Silver-tier
SHARED			
Cluster size (GPUs)	5184	5,184	5,184
Avg job size (j_size) (GPUs)	4096	4,096	4,096
Blast radius (b_radius) (GPUs)	64	64	64
PER PROVIDER			
GPU MTBF (GPU-hrs)	25000	25000	15000
Checkpoint freq (t_chkpt) (mins)	60	60	60
Failover time (t_failover) (mins)	5	5	5
Idle spare GPUs (GPUs)	32	0	0
RESILIENCY	Fault Tolerant	Fault Tolerant	Checkpoint Restart
Repair/Replace	TorchPass	HyperPod Ckptless	Hot spare (idle)
Time to identify failure (mins)	15	15	60
Time to repair node (t_repair) (hrs)	0.25	0.25	1
Time to init job (mins)	10	10	15
Network overhead (%)	0	0	0
Memory overhead (%)	0	5	0
RESULTS			
Cluster MTBF	4.8h	4.8h	2.9h
Interruptions/mo	149.3	149.3	248.8
Downtime loss	5.53%	5.53%	20.91%
Idle spare cost	0.62%	—	—
Performance overhead	0.00%	5.00%	0.00%
Total Goodput Loss	**6.14%**	**10.53%**	20.91%

Source: https://clustermax.ai/tco

INFERENCEX

Open-source continuous inference benchmark — the bridge between raw hardware and $/Mtok

WHAT IT MEASURES

- Interactivity (tok/s/user) and per-GPU throughput (tok/s/GPU)
- P99 time-to-first-token (TTFT) across concurrency levels
- Tokens per megawatt (tok/s/MW) and joules per token
- $ per million tokens at hyperscaler, neocloud, and rental price

Every datapoint is a public GitHub Actions run — click any chart point to inspect the recipe, logs, and artifacts. Weekly DB releases for audit.

COVERAGE

NVIDIA	H100, H200, B200, B300, GB200, GB300
AMD	MI300X, MI325X, MI355X
Models	DeepSeek-R1, gpt-oss-120B, Llama-3.3-70B, Qwen, Kimi-K2, GLM-5, MiniMax-M2
Frameworks	vLLM, SGLang, TRT-LLM, Dynamo (TRT/vLLM/SGLang)
Precisions	FP4, FP8, BF16, INT4
Modes	Disaggregated prefill/decode, MTP, wide expert parallelism for MoE

Source: InferenceX (formerly InferenceMAX), InferenceX v2: Blackwell vs AMD vs Hopper

FROM BENCHMARK TO ECONOMICS

InferenceX gives you tok/s/GPU and joules/token. Add price and SLA and you have $/Mtok — the unit the next section is built on.

1

BENCHMARK

InferenceX measures tok/s/GPU, P99 TTFT, joules/token on real hardware across all models, frameworks, precisions.

2

PRICE

Apply the GPU rental price (from ClusterMAX-rated providers) or hyperscaler list price to the throughput.

3

$/MTOK

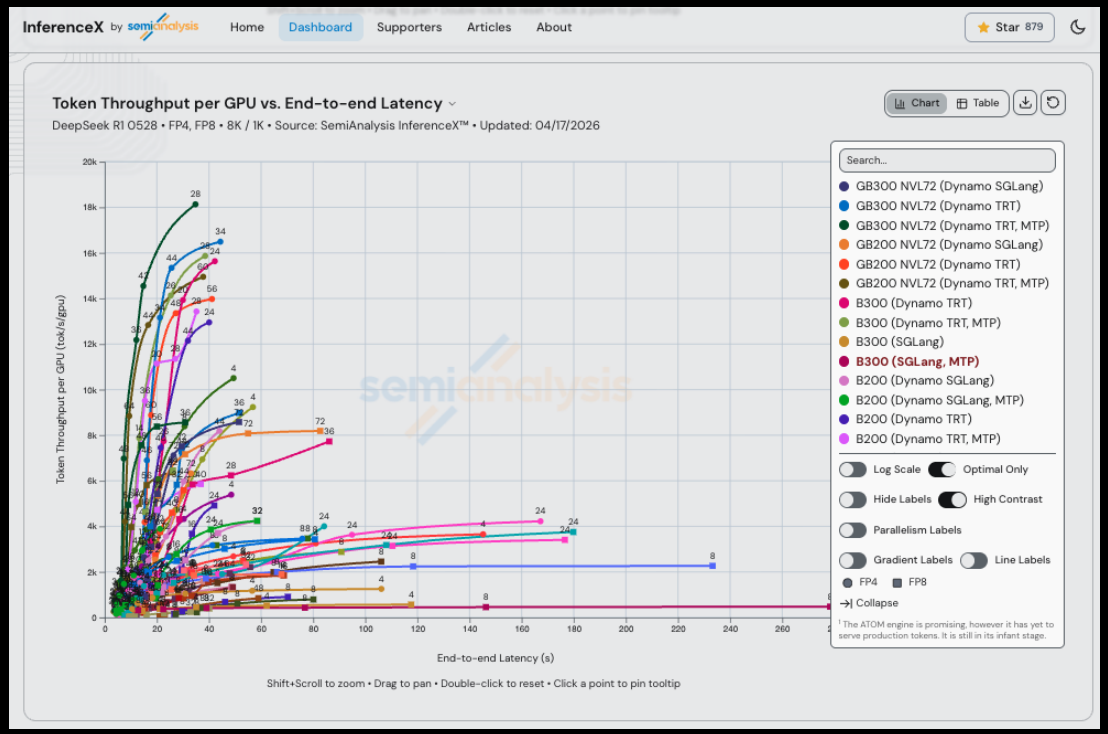
Result: cost per million tokens at your SLA — the only unit the customer actually feels and pays for.

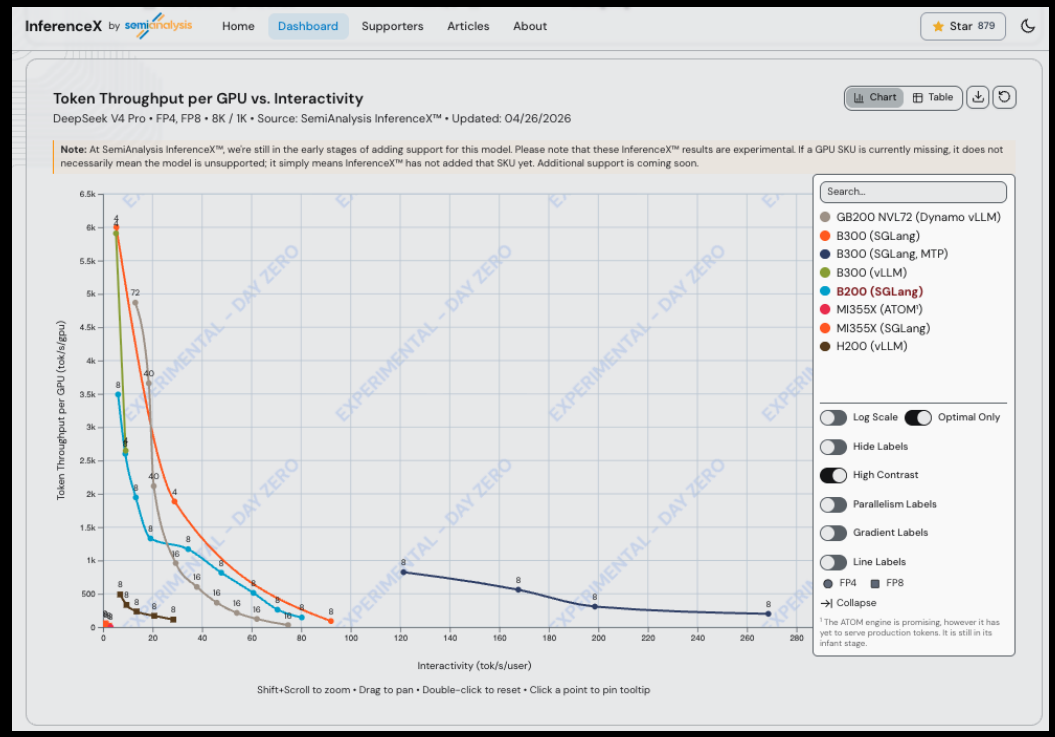
Up next — tokenomics: who pays the $/Mtok

Source: InferenceX, ClusterMAX

INTRODUCTION TO INFERENCEX

An **open-source, automated benchmark** that moves at the same speed as the AI software ecosystem



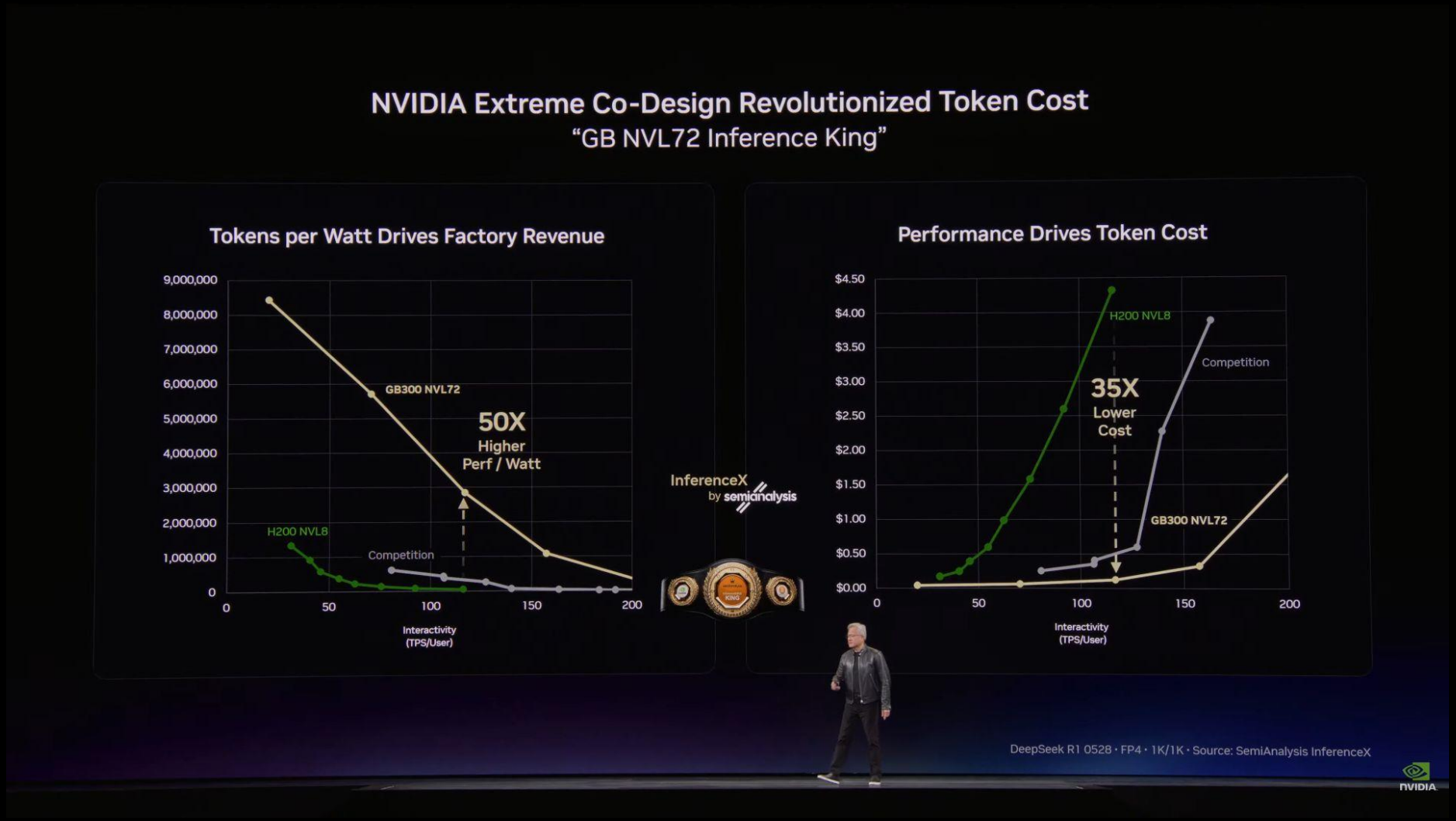


featuring day zero support for DeepSeek V4!

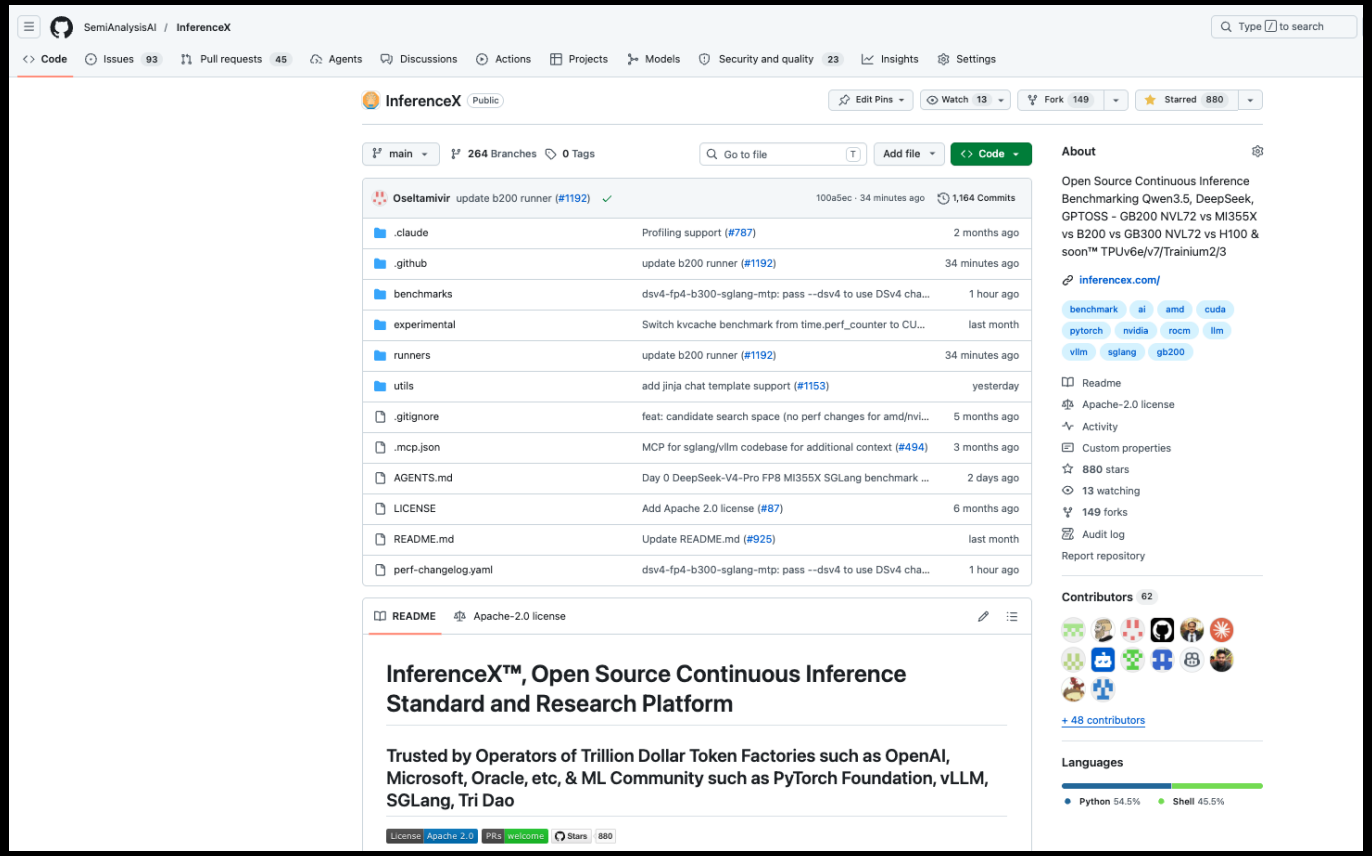
Source: inferencex.com

INTRODUCTION TO INFERENCEX

Featured on stage at the GTC 2026 Keynote by Inference King Jensen Huang



UNDERSTANDING INFERENCEX



Software evolves every single day, delivering **continuous performance gains** on top of step jumps in hardware

AI software like SGLang, vLLM, TensorRT-LLM, CUDA and ROCm increase the **pareto frontier of performance** in incremental releases that are just **days apart**

Benchmarks conducted at a fixed point in time quickly go stale and are not representative of what can be achieved with the latest software packages

Source: github.com/SemiAnalysisAI/inferencex

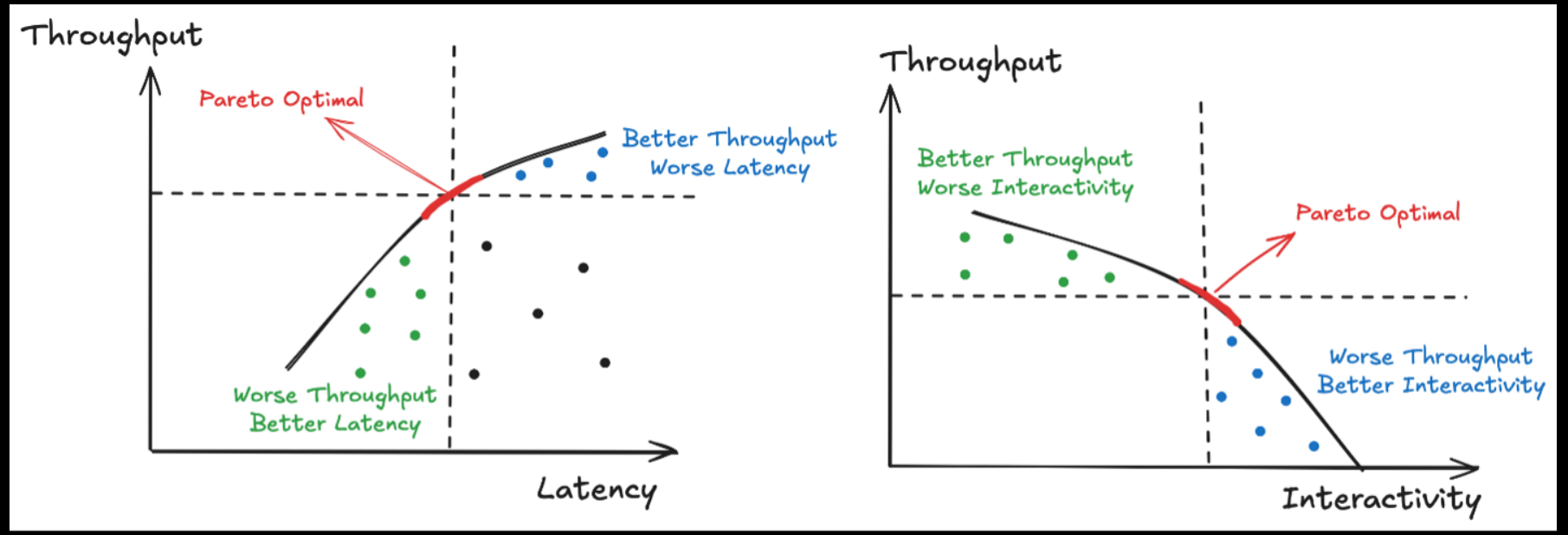
UNDERSTANDING INFERENCEX: THE PARETO FRONTIER

What It Is

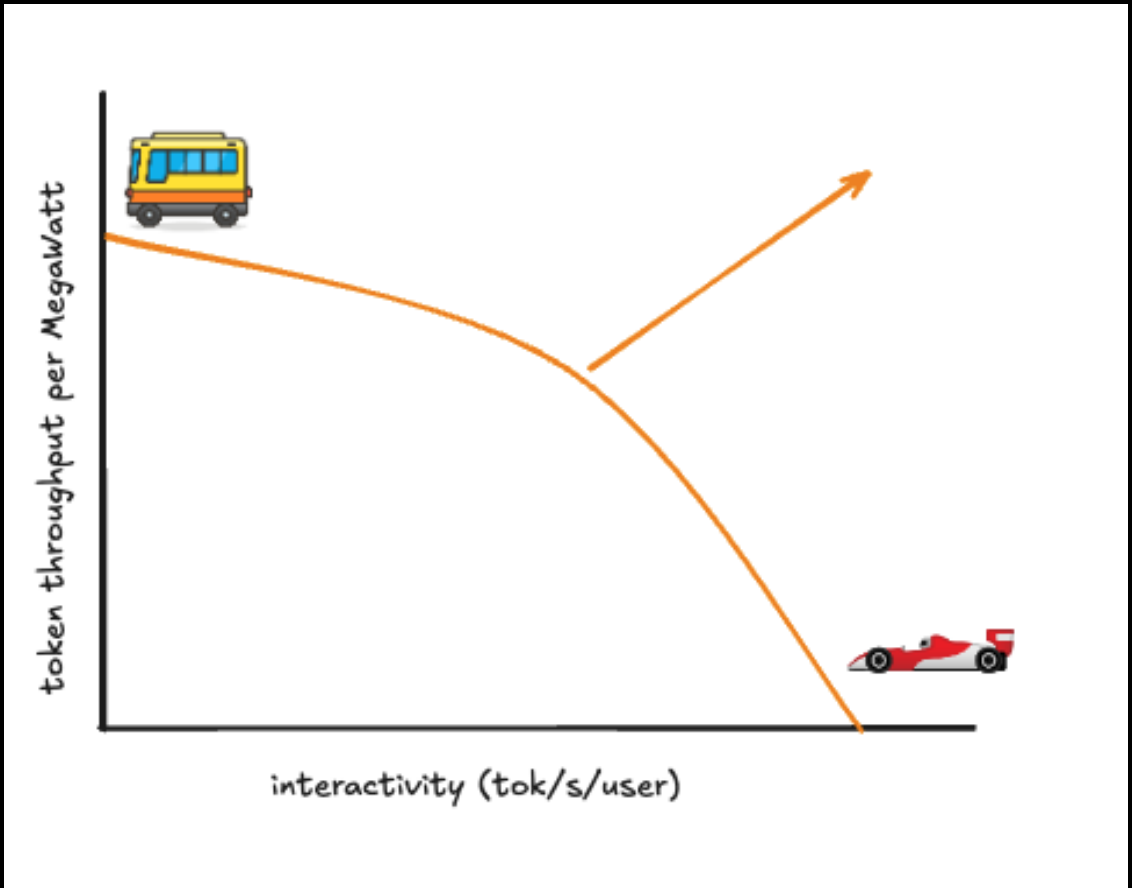
- Curve showing **best possible** throughput vs latency trade-offs
- **Pareto optimal point**: No other point improves one metric without sacrificing the other
- Points below frontier = **suboptimal**, wasting performance

Why It Matters

- Operating below frontier = wasted resources = **lost revenue**
- Moving to frontier = **serve more users** OR **deliver faster responses** with same infrastructure
- Frontier = **maximum revenue potential**



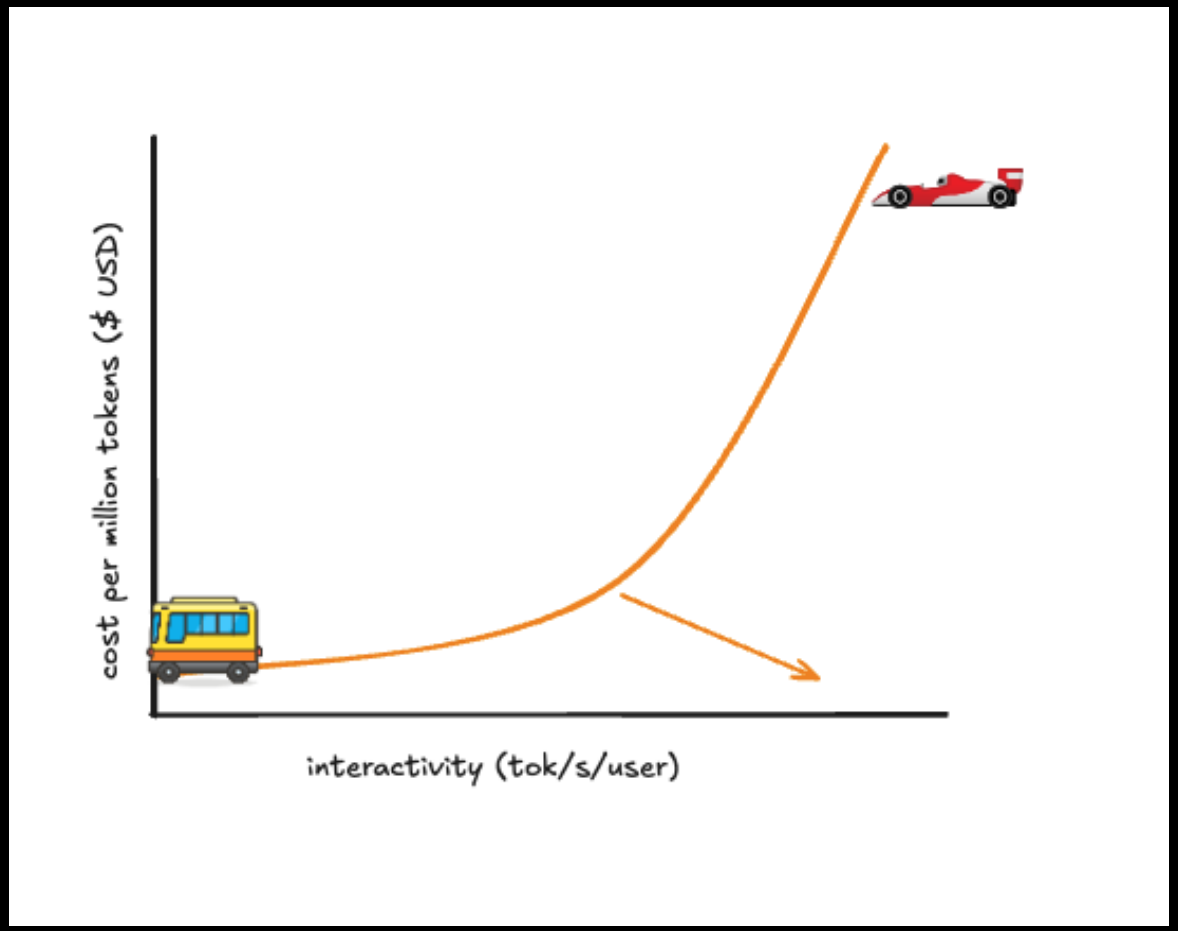
TOKEN THROUGHPUT VS INTERACTIVITY (TOK/S/USER)



Why It Matters

- Electricity & colocation cost make up to **20% of TCO**
- 20% lower tok/s/MW will only impact **4% of TCO**
- Higher interactivity → fewer tokens per datacenter
- Lower interactivity → higher tokens per datacenter

TCO PER MILLION TOKENS VS INTERACTIVITY (TOK/S/USER)



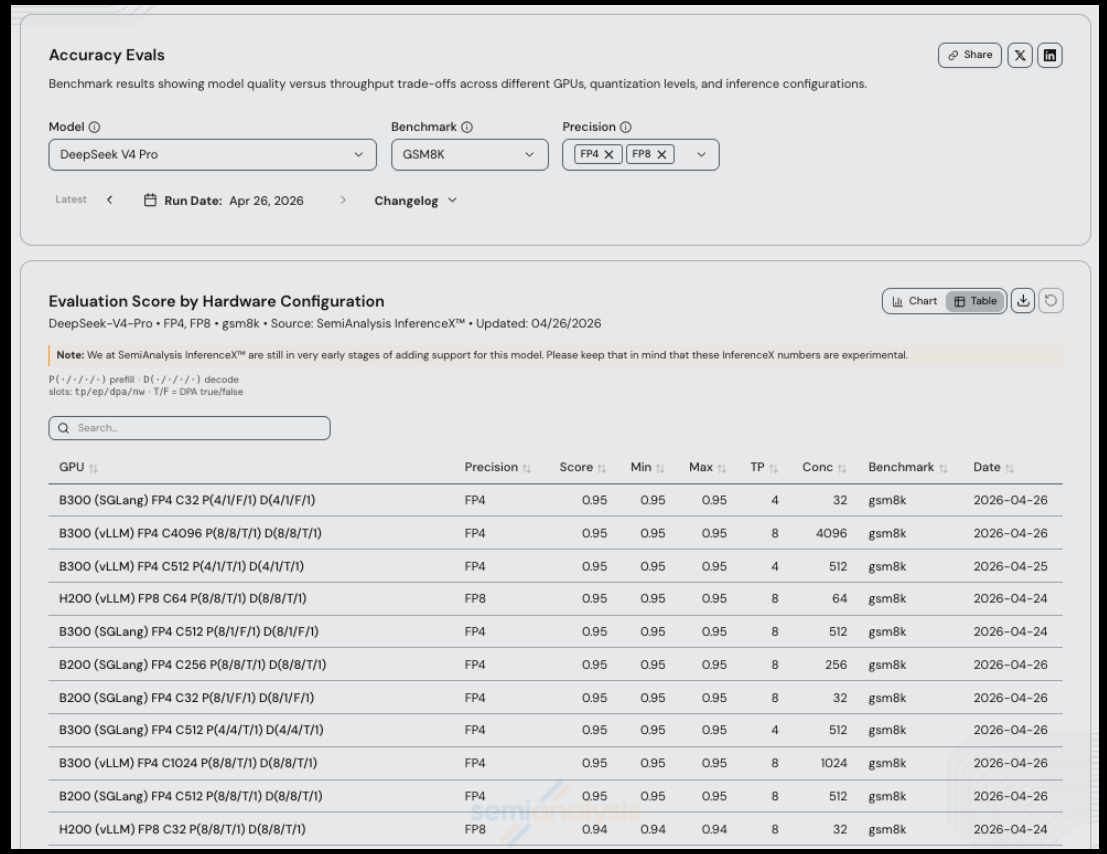
Why It Matters

- Higher interactivity → fewer tokens/hour → higher $/token
- Higher throughput → more tokens/hour → lower $/token
- **Premium responsiveness** requires **premium pricing**

ADDITIONAL FEATURES: GOING BEYOND PERFORMANCE

Accuracy Evals

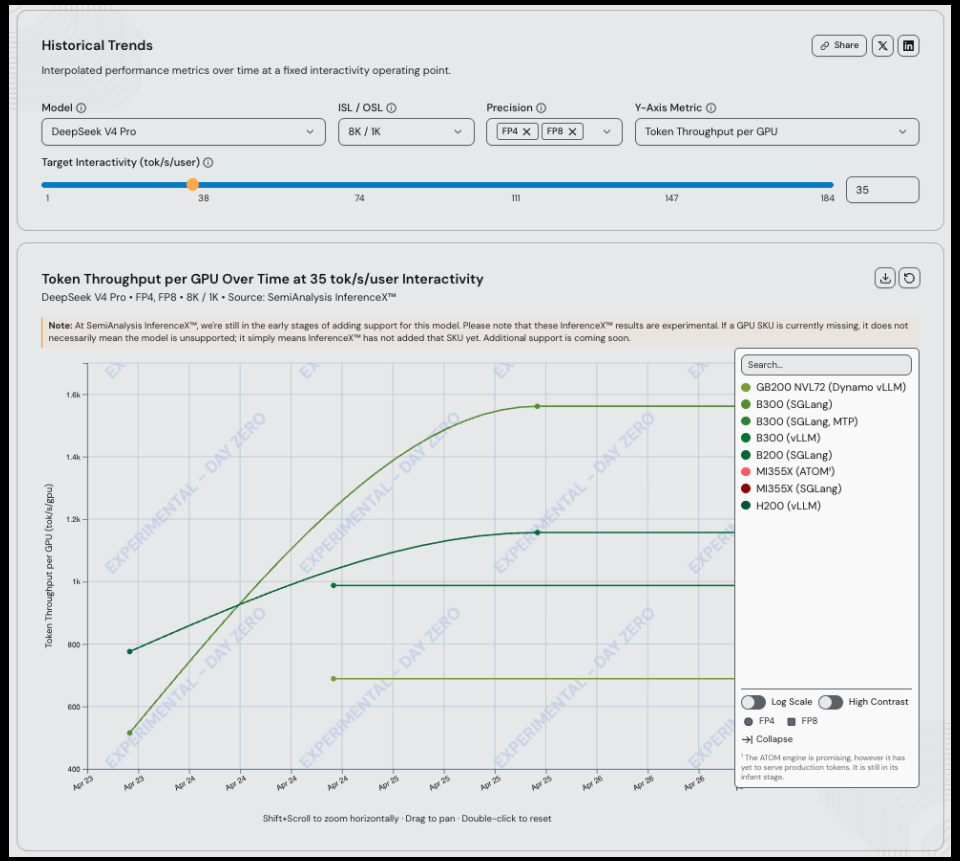
Confirming that performance improvements do not sacrifice model quality



GPU	Precision	Score	Min	Max	TP	Conc	Benchmark	Date
B300 (SGLang) FP4 C32 P(4/1/F/1) D(4/1/F/1)	FP4	0.95	0.95	0.95	4	32	gsm8k	2026-04-26
B300 (vLLM) FP4 C4096 P(8/8/T/1) D(8/8/T/1)	FP4	0.95	0.95	0.95	8	4096	gsm8k	2026-04-26
B300 (vLLM) FP4 C512 P(4/1/T/1) D(4/1/T/1)	FP4	0.95	0.95	0.95	4	512	gsm8k	2026-04-25
H200 (vLLM) FP8 C64 P(8/8/T/1) D(8/8/T/1)	FP8	0.95	0.95	0.95	8	64	gsm8k	2026-04-24
B300 (SGLang) FP4 C512 P(8/1/F/1) D(8/1/F/1)	FP4	0.95	0.95	0.95	8	512	gsm8k	2026-04-24
B200 (SGLang) FP4 C256 P(8/8/T/1) D(8/8/T/1)	FP4	0.95	0.95	0.95	8	256	gsm8k	2026-04-26
B200 (SGLang) FP4 C32 P(8/1/F/1) D(8/1/F/1)	FP4	0.95	0.95	0.95	8	32	gsm8k	2026-04-26
B300 (SGLang) FP4 C512 P(4/4/T/1) D(4/4/T/1)	FP4	0.95	0.95	0.95	4	512	gsm8k	2026-04-26
B300 (vLLM) FP4 C1024 P(8/8/T/1) D(8/8/T/1)	FP4	0.95	0.95	0.95	8	1024	gsm8k	2026-04-26
B200 (SGLang) FP4 C512 P(8/8/T/1) D(8/8/T/1)	FP4	0.95	0.95	0.95	8	512	gsm8k	2026-04-26
H200 (vLLM) FP8 C32 P(8/8/T/1) D(8/8/T/1)	FP8	0.94	0.94	0.94	8	32	gsm8k	2026-04-24

Historical Trends

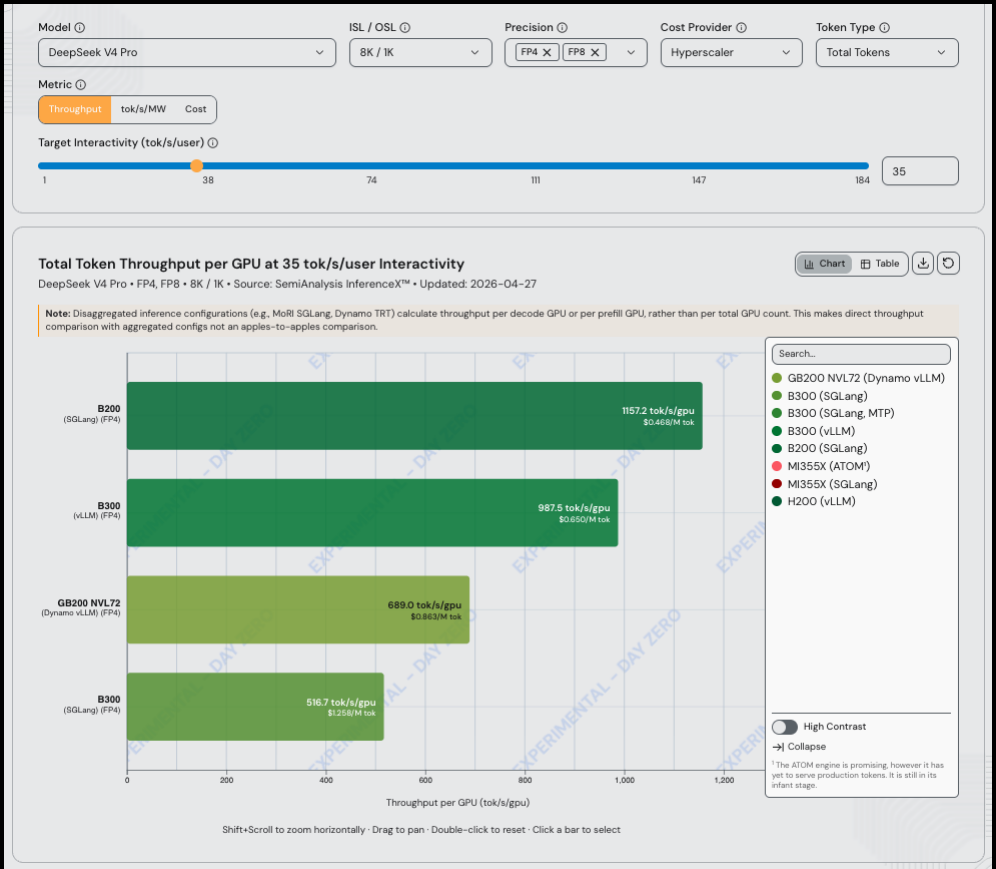
At iso-interactivity, demonstrating throughput improvements over time



ADDITIONAL FEATURES: GOING BEYOND PERFORMANCE

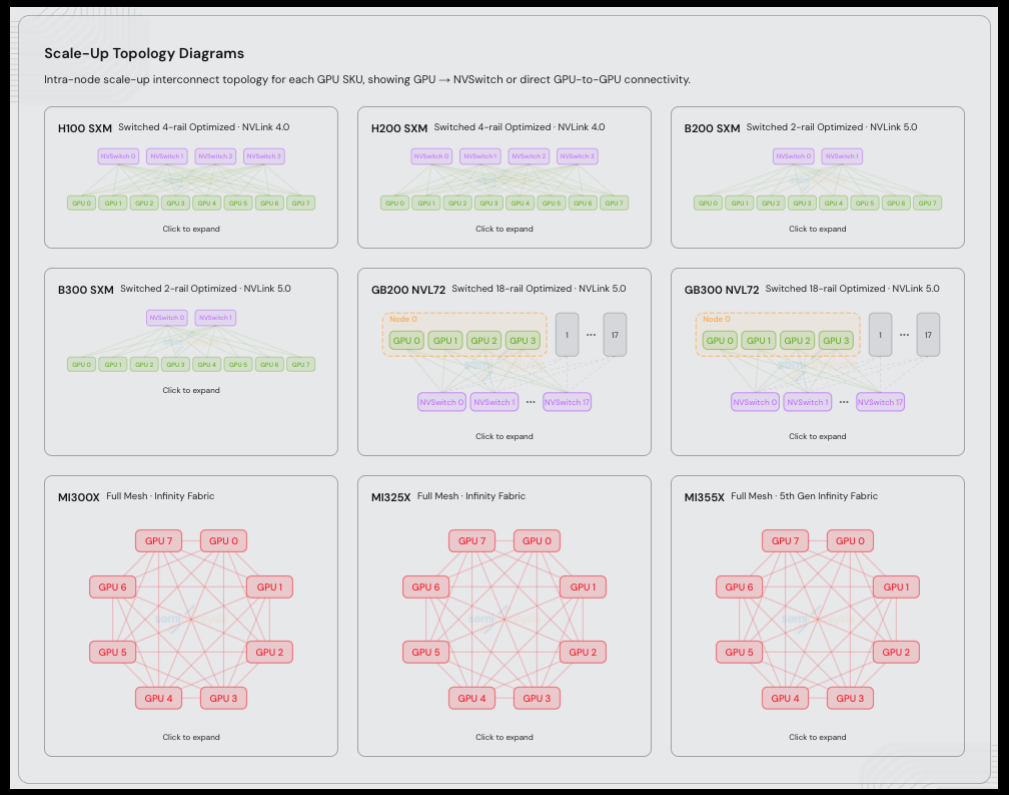
TCO Calculator

Assessing which hardware is best for a given model



Interactive Diagrams

Understanding GPU Systems, Model Architectures and More



TOKENOMICS: WHERE DOES THE VALUE GO?

The AI Token Factory Economics Stack				
Layer	**Units**	**H200**	**Business**	**SemiAnalysis Model**
Total Server Capital Cost	***USD per 8 GPU***	**$246,391**	Chip	Accelerator and AI TCO Model
Useful Life	*Yrs*	4.0	IAAS	AI TCO Model
Depreciation Expense/GPU	***USD/hr/GPU***	**$0.88**	IAAS	
Colocation Cost	*USD/kW/mth*	$110	IAAS	AI Datacenter Model
Electricity Cost	*USD/kWh*	$0.087	IAAS	
Operating Cost per GPU	***USD/hr/GPU***	**$0.35**	IAAS	
Total Variable Cost per GPU	***USD/hr/GPU***	**$1.23**	IAAS	AI TCO Model
Revenue per GPU	***USD/hr/GPU***	**$1.90**	IAAS	
Neocloud Gross Margin	*%*	*35%*	IAAS	
Effective Inference Throughput/GPU[1]	*Tokens/s/GPU*	735	PAAS	InferenceMAX™
Inference Compute Cost	***USD/M Tok***	**$0.72**	PAAS	
Inference API Blended Price	***USD/M Tok***	**$1.75**	PAAS	
Model Gross Margin	*%*	*59.0%*	PAAS	
Tokens/User/Month	*Tok/user/mth*	10,000,000	Application	AI Tokenomics Model
3P Token Unit Cost	*USD/M Tok*	$1.75	Application	
Average Revenue per User	*USD/mth*	$20.00	Application	
3P Gross profit per User/Mth	***USD/mth***	**$2.50**	Application	
3P Application Gross Margin	*%*	*12.5%*	Application	
1P Token Unit Cost	*USD/M Tok*	$0.72		AI Tokenomics Model
1P Gross profit per User/Mth	***USD/mth***	**$12.82**	Application	
1P Application Gross Margin	*%*	*64.1%*	Application	

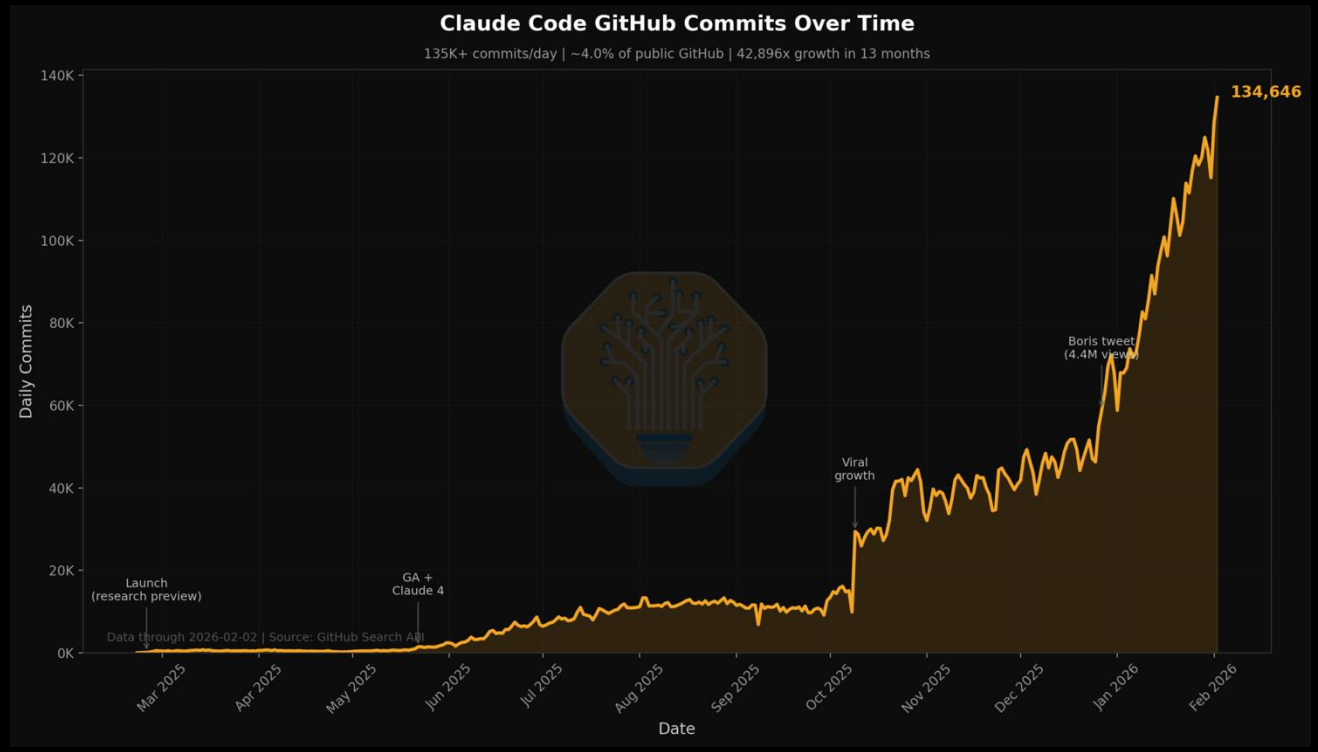
1. DeepSeek R1 FP8 on H200 using SGLang. Uses 8k input, 1k output tokens, 43 interactivity.

PART 04

End User Demand

Tokenomics: who actually uses the tokens, how it is monetized, and whether the revenue justifies the buildout.

CLAUDE CODE HAS ARRIVED



semianalysis

Claude Code is the Inflection Point

What It Is, How We Use It, Industry Repercussions, Microsoft's Dilemma, Why Anthropic Is Winning

DOUG O'LAUGHLIN, JEREMIE ELIAHOU ONTIVEROS, JORDAN NANOS, AND 2 OTHERS

FEB 05, 2026 • PAID

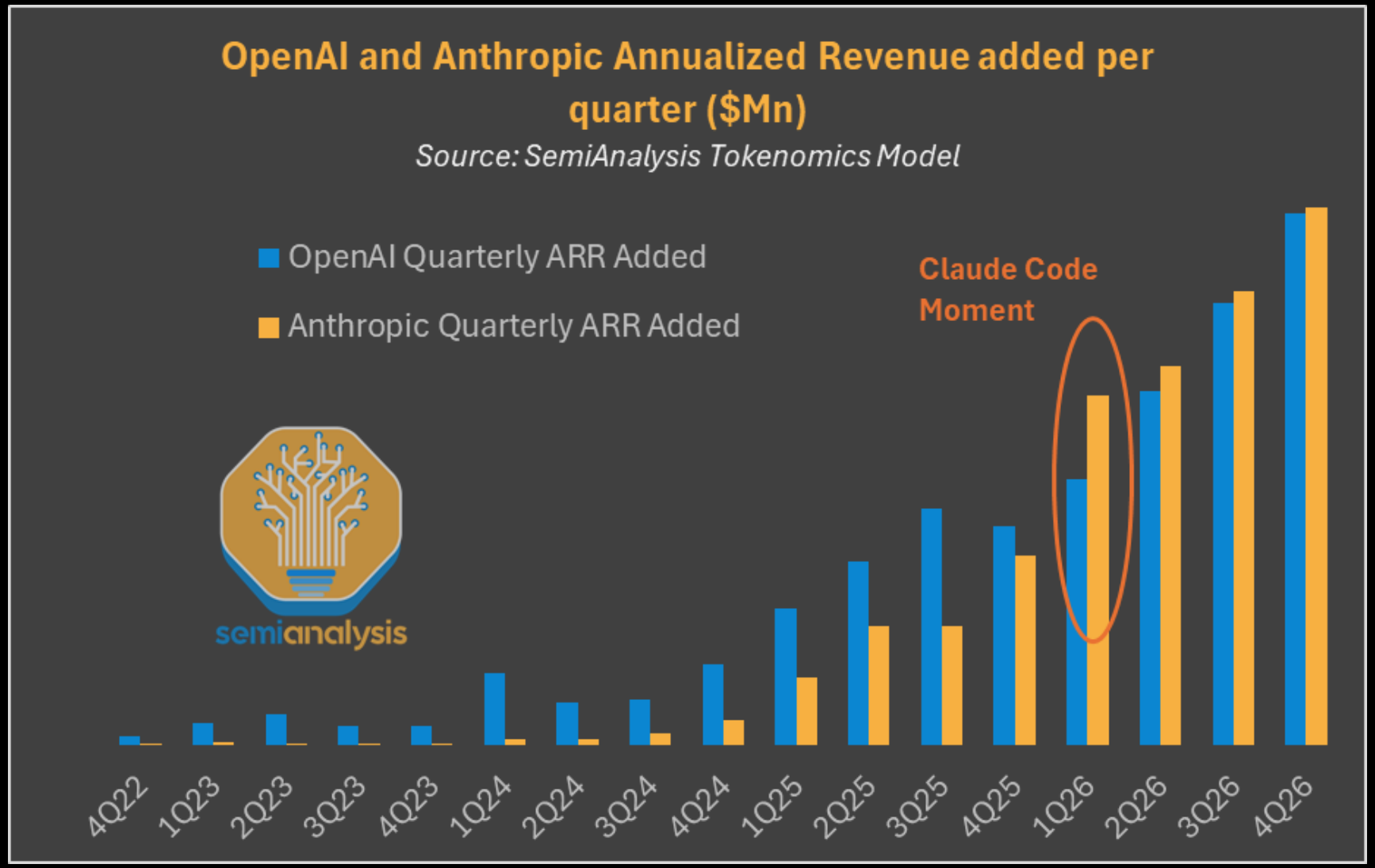
538 6 81 Share

4% of GitHub public commits are being authored by Claude Code right now. At the current trajectory, we believe that Claude Code will be 20%+ of all daily commits by the end of 2026. While you blinked, AI consumed all of software development.

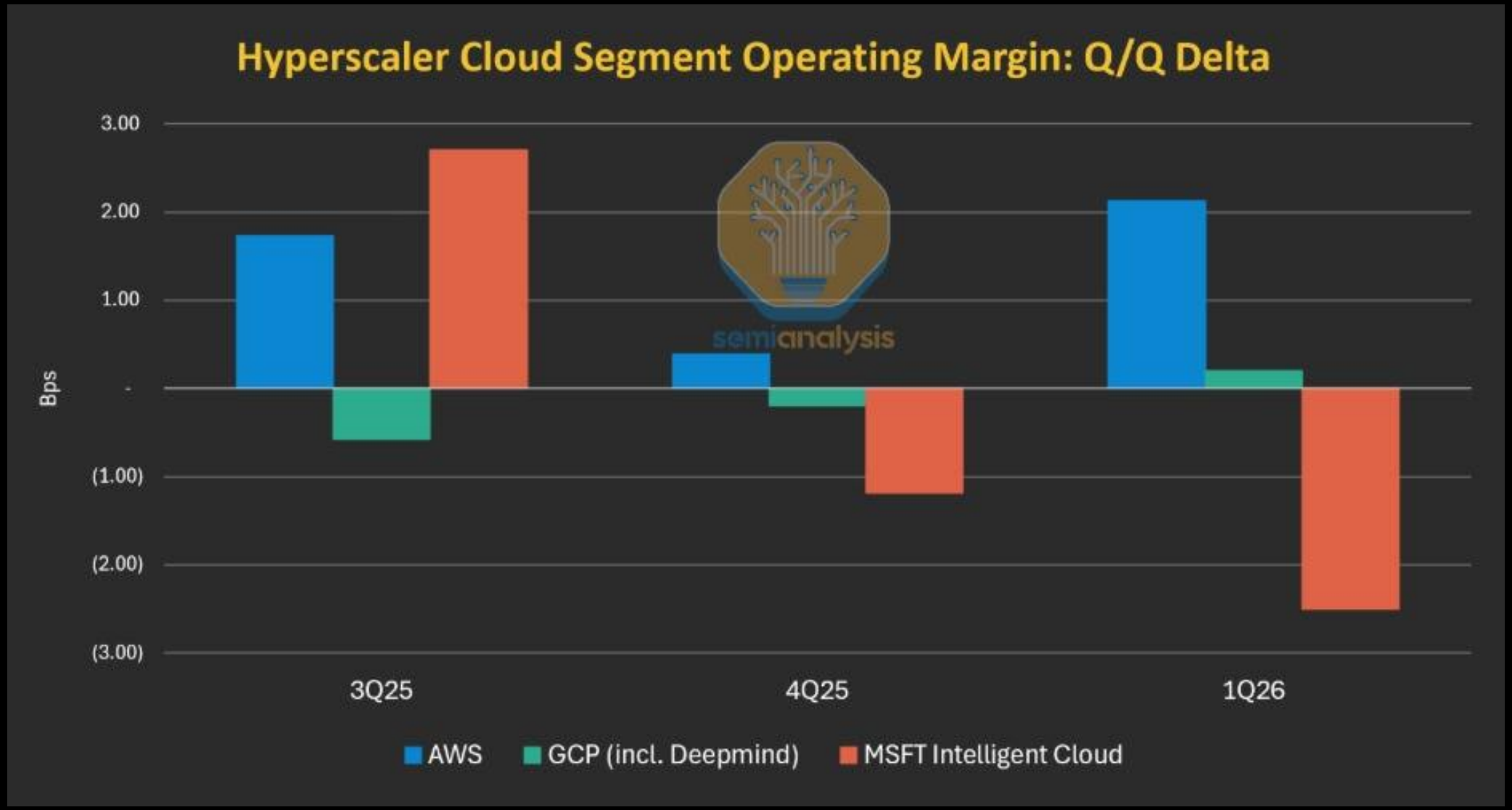
Our sister publication Fabricated Knowledge described software like linear TV during the rise of the internet and thinks that the rise of Claude Code is going to be a new layer of intelligence on top of software akin to DRAM versus NAND. Today SemiAnalysis is going to dive into the repercussions of Claude Code, what it is, and why Claude is so good.

Source: https://newsletter.semianalysis.com/p/claude-code-is-the-inflection-point

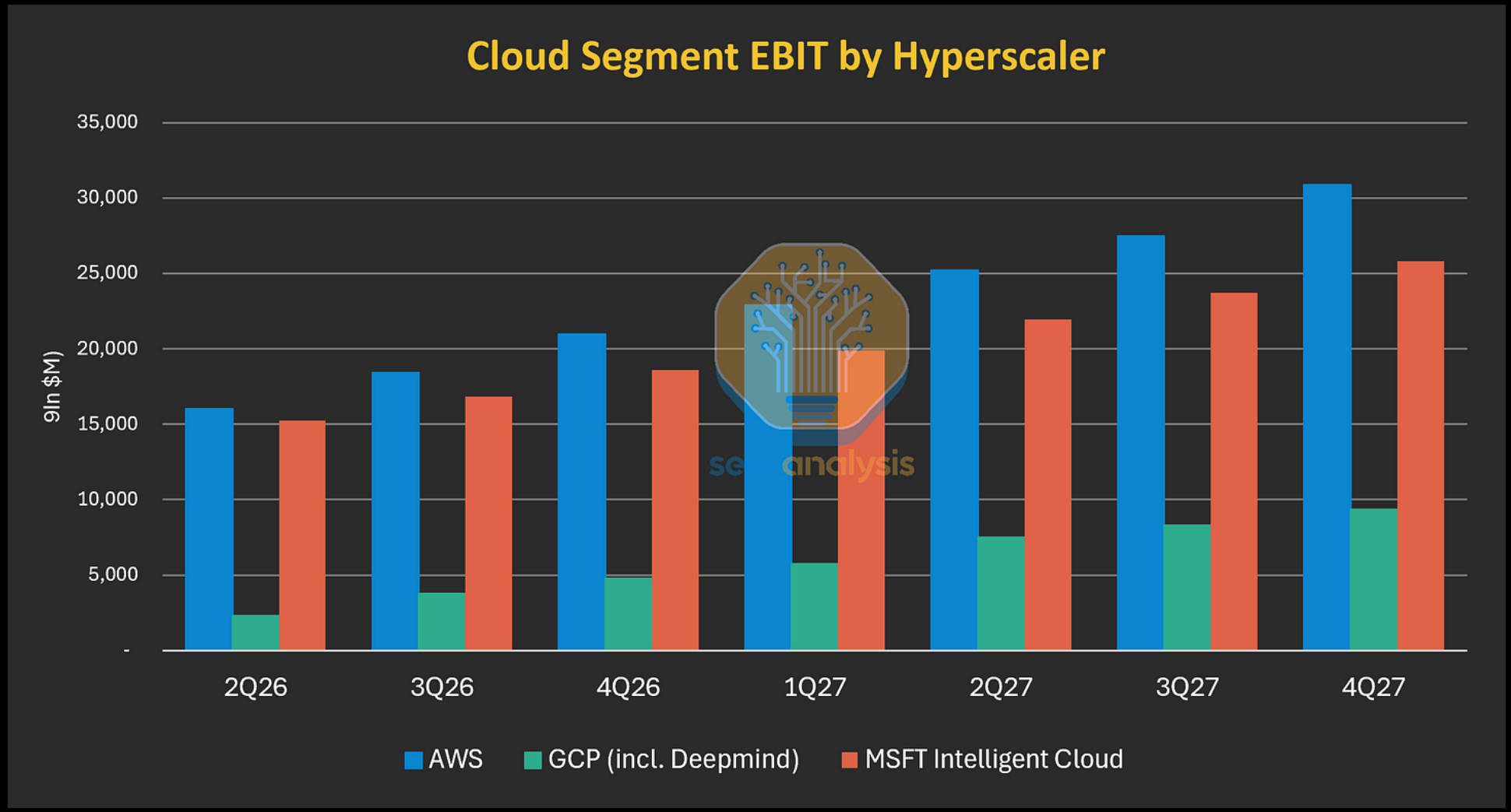
ANTHROPIC GROWTH AS A RESULT



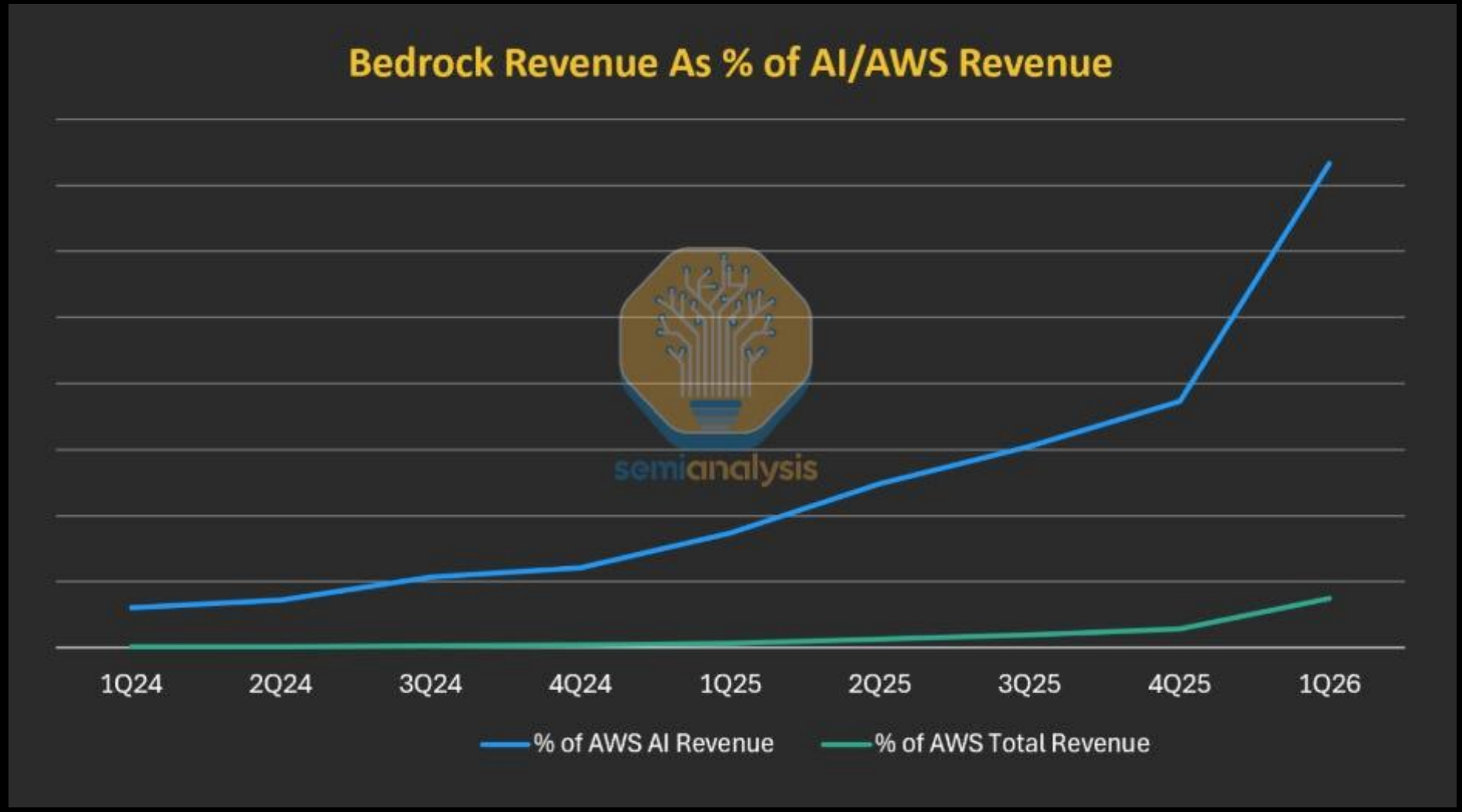
AMAZON IS A KEY BENEFICIARY OF THIS GROWTH



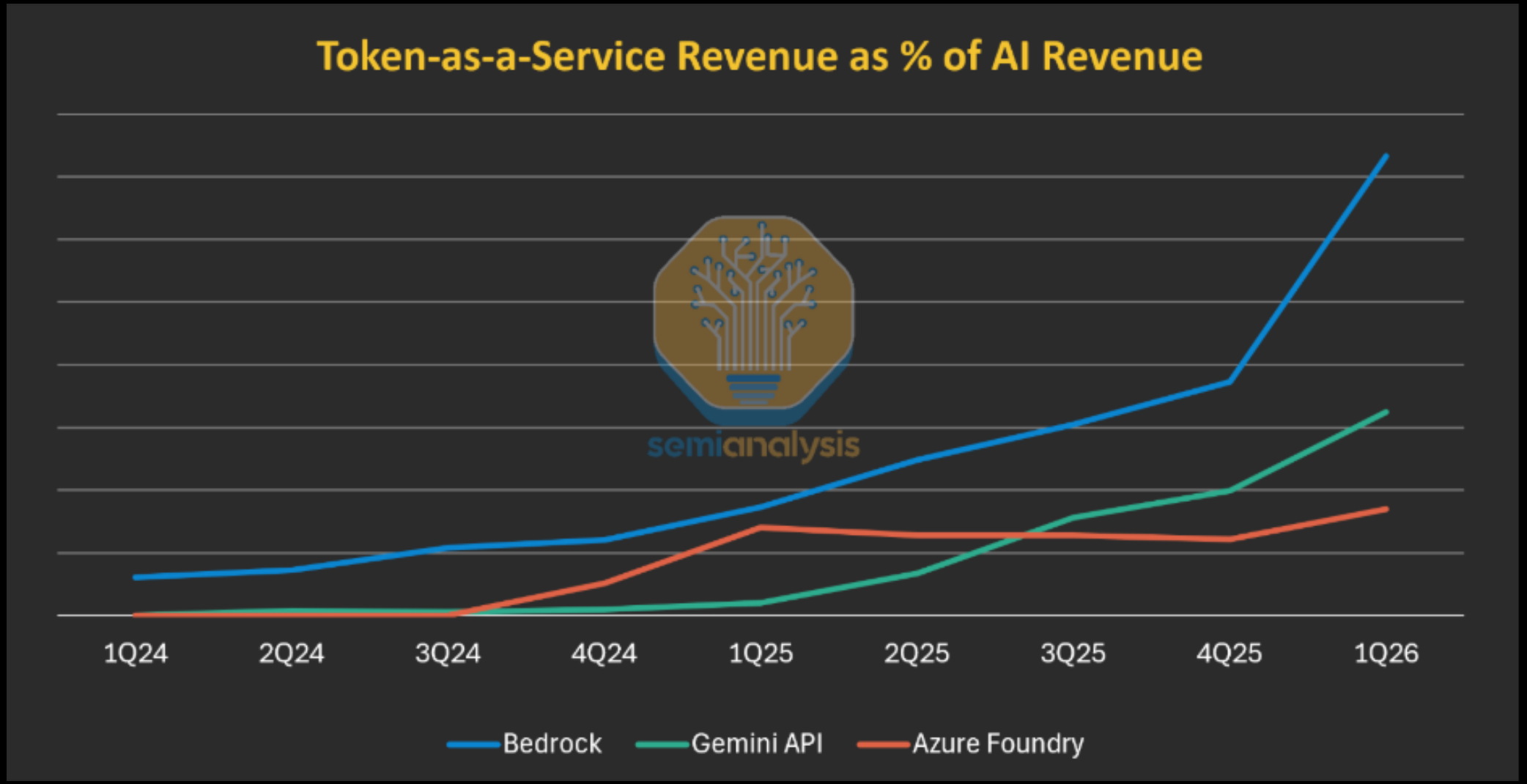
AMAZON IS A KEY BENEFICIARY OF THIS GROWTH



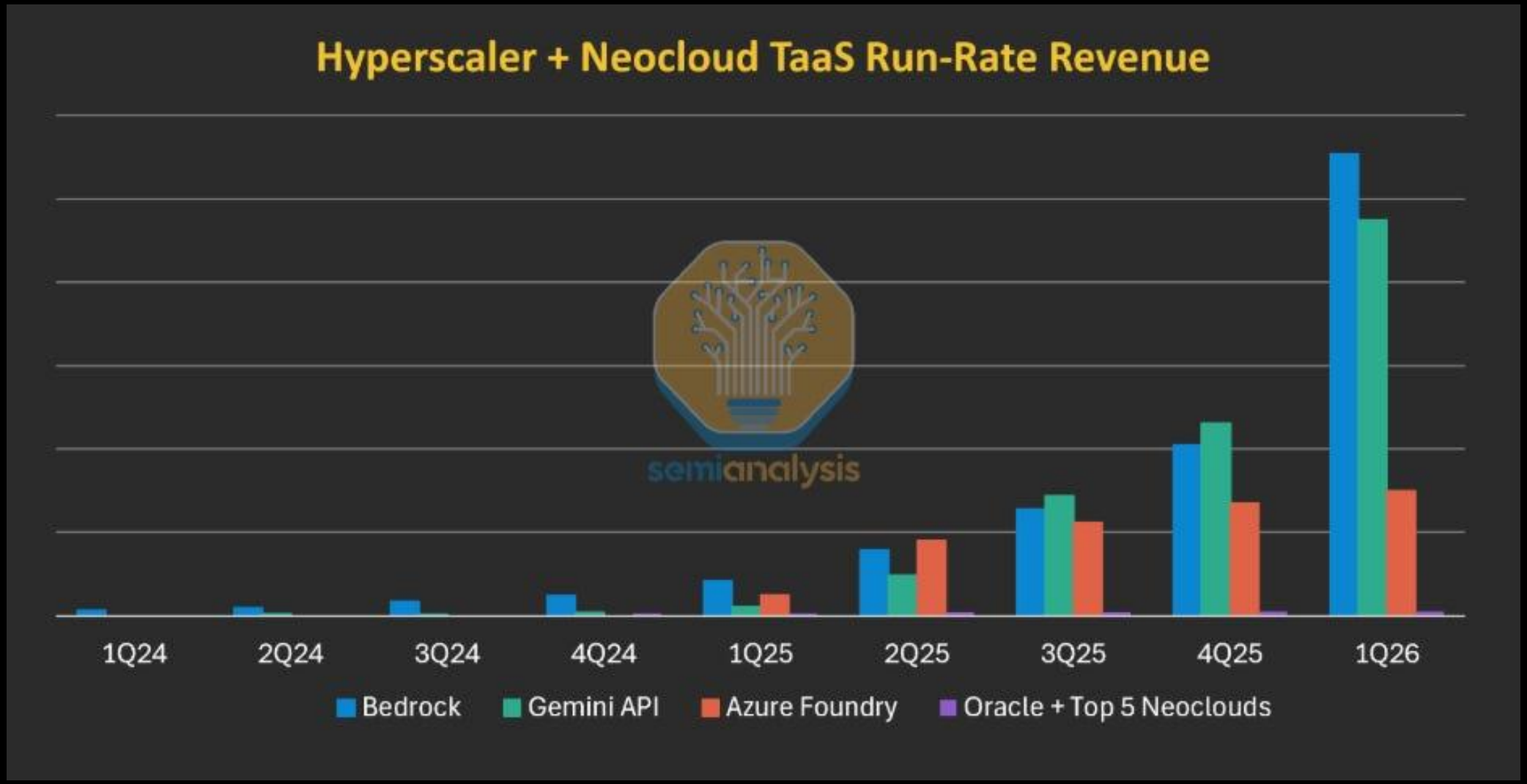
BEDROCK IS GROWING LIKE CRAZY



TOKEN GROWTH IS STRONG



NEOCLOUDS ARE LARGELY IRRELEVANT



WE EXPECT THIS TO CONTINUE

LOOK AT HOW FAST THEY'RE BUILDING

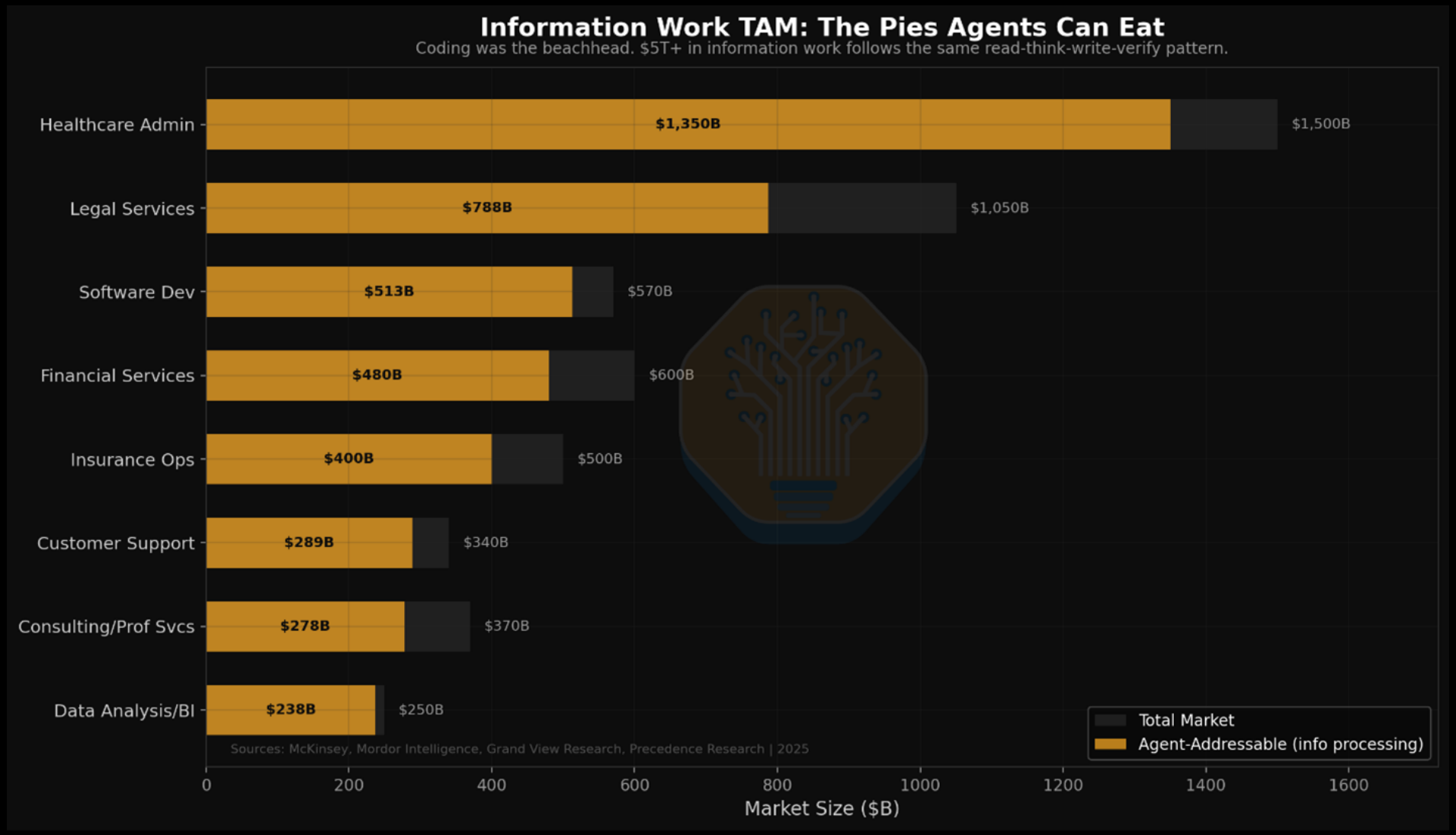
AWS / Anthropic AI Training Datacenters as of **Q3 2025**

Source: SemiAnalysis Datacenter Industry Model

AWS / Anthropic AI Training Datacenters as of **Q3 2024**

Source: SemiAnalysis Datacenter Industry Model

AGENTS ARE JUST GETTING STARTED



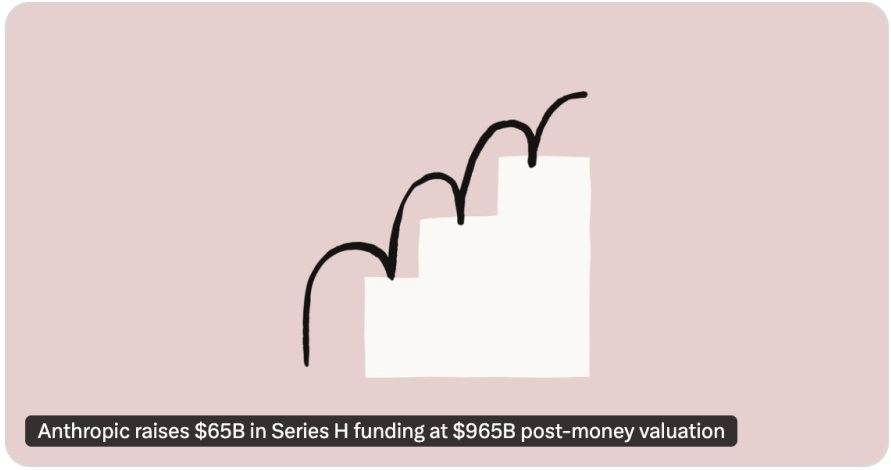
OPENAI & ANTHROPIC ARE HIGHLY PROFITABLE

Anthropic ✓
@AnthropicAI

Earlier this month, our run-rate revenue crossed $47 billion.

This growth has been driven by organizations across many industries deploying Claude in their core operations, and by a growing number of people using it for their everyday work.

Read more:



From anthropic.com

source: https://www.anthropic.com/news/series-h

OpenAI was the fastest technology platform to reach 10 million users, the fastest to 100 million users, and soon the fastest to 1 billion weekly active users. Within a year of launching ChatGPT, we reached $1B in revenue. By the end of 2024 we were generating $1B per quarter. We are now generating $2B in revenue per month. At this stage, we are growing revenue four times faster than the companies who defined the Internet and mobile eras, including Alphabet and Meta.

source: https://openai.com/index/accelerating-the-next-phase-ai/

← **Post**

Joey Brookhart ✓
@SaasquatchC

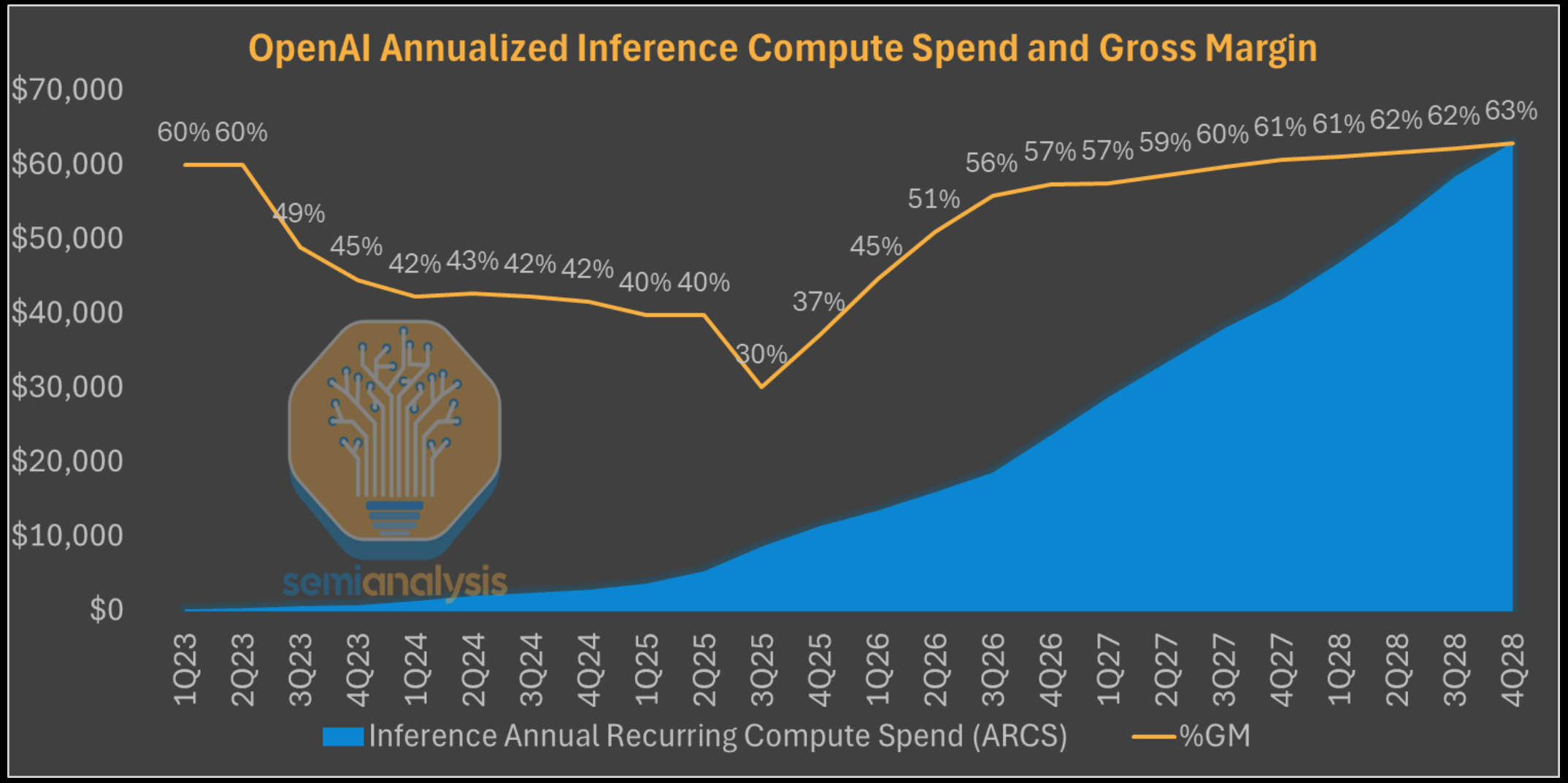
Congrats to our Tokenomics team at SemiAnalysis who had the Anthropic revenue numbers, cost buckets, and adjusted EBIT numbers dead on for Q1 and Q2

7:40 PM · May 20, 2026 · **20.3K** Views

2 6 205 47

Relevant ⌄ View quotes ›

THE "SHIFT FROM TRAINING TO INFERENCE" IS A LIE



THANK YOU

QUESTIONS?

JORDAN NANOS

Member of Technical Staff · SemiAnalysis

Newsletter	newsletter.semianalysis.com
SemiAnalysis	semianalysis.com
ClusterMAX	clustermax.ai
InferenceX	inferencex.com
Tokenomics	tokenomics.info

THE TAKEAWAY

- Supply chain (CoWoS, HBM) decides who gets chips.
- Power + cooling innovations defines where they land.
- Networking decides how they perform.
- Tokenomics decides who earns the ROIC.

semianalysis